The validity of the practical examination from the perspective of experts in initial vocational education in the retail trade sector in Switzerland[1]

La validité de l'examen pratique sous le regard des experts en formation professionnelle initiale dans la branche du commerce de détail en Suisse

A validade do exame prático sob a perspectiva de especialistas em formação profissional inicial no setor de comércio varejista na Suíça

David JAN
ID ORCID: 0000-0001-5223-1685
Université de Fribourg

Keywords: vocational education, latent classes, assessment, certificate exams, examiner

In Switzerland, at the end of their initial vocational education, retail apprentices take a practical exam at the company where they are being trained. This exam, which is both certifying and eliminating, lasts between one and one and a half hours, depending on the type of training. It is conducted by two professionals in the field known as examiners. The objectives of this study are, firstly, to understand the validity attributed to the practical exam by the examiners and, secondly, to define the reasons for this assessment. To achieve this, a sample of 189 examiners answered a questionnaire on the validity of this exam. The data were processed by exploratory factor analysis, and three expert profiles were extracted by latent classes. The results of this study highlighted the heterogeneity of the examiners profiles. They differ according to their experience as examiners and their language.

1. The French version was published in issue 47(2) 2024: https://doi.org/10.7202/1117466ar

Mots clés : classes latentes, évaluateur, évaluation, examen certificatif, formation professionnelle initiale

En Suisse, en formation professionnelle initiale dans la branche du commerce de détail, la diplomation passe par l'examen pratique dans l'entreprise formatrice de l'apprenti. Cet examen, certificatif et éliminatoire, dure entre une heure et une heure et demie, selon le type de formation. Il est conduit par deux professionnels du domaine nommés experts. Les buts de cette étude sont, d'une part, de comprendre la validité attribuée par les experts à l'examen pratique et, d'autre part, de définir les raisons de cette évaluation. Pour y parvenir, un échantillon de 189 experts a répondu à un questionnaire sur la validité de cet examen. Les données ont été traitées par des analyses factorielles exploratoires. Nous avons ensuite extrait trois profils d'experts par classes latentes. Les résultats de cette étude ont permis de mettre en évidence, parmi les experts, une hétérogénéité de profils. Ils se différencient selon leur expérience en tant qu'expert et selon leur langue.

Palavras-chaves: avaliação, avaliador, classes latentes, exame certificativo, formação profissional inicial

Na Suíça, na formação profissional inicial no setor do comércio a retalho, a atribuição de diplomas é feita através de um exame prático realizado na empresa formadora do aprendiz. Este exame, de caráter certificativo e eliminatório, tem uma duração de uma hora a uma hora e meia, dependendo do tipo de formação. Ele é conduzido por dois profissionais da área, designados como peritos. Os objetivos deste estudo são, por um lado, compreender a validade atribuída pelos peritos ao exame prático e, por outro, identificar as razões dessa avaliação. Para tal, foi utilizado um questionário sobre a validade do exame, respondido por uma amostra de 189 peritos. Os dados foram analisados por meio de análises fatoriais exploratórias, a partir das quais foram extraídos três perfis de peritos através de classes latentes. Os resultados deste estudo evidenciaram uma heterogeneidade de perfis entre os peritos, diferenciando-se pela sua experiência enquanto peritos e pela sua língua.

Author's note: Correspondence concerning this article can be addressed to david.jan@unifr.ch

Introduction

Switzerland has a well-established dual system of initial vocational education and training (VET), where apprentices sign a contract with a host company for employment and training two to three days a week. The rest of their time is broken down between attending vocational school and inter-company courses according to State Secretariat for Education, Research and Innovation (SERI) directives (2022). The aim is to provide apprentices with training that is pertinent to the labor market, so they will be ready to practice their profession independently (SERI, 2022). Skill assessment is ongoing and at the end of the course, when the apprentices undertake various exams, including a single practical exam. This summative exam requires a sufficient mark for graduation. The practical exam, which takes place in the apprentice's training company, lasts for 60 or 90 minutes depending on the type of training, followed by a 30-minute marking session. The practical exam must be conducted by two examiners who are active professionals in the trade and from another company. The examiners for this practical exam are known as "company examiners"; in everyday language, only the term "examiner" is used.

This important dual training program has been extensively documented in scientific research. Studies have focused on the effectiveness of organization, analyzing the reasons for the 20–25% premature termination of apprenticeship contracts (Kriesi et al., 2016; FSO, 2023; Schmid & Stalder E., 2008); coordination between the various training centers (Berger et al., 2021, 2023; Frey, 2010); and the professional future of graduates (Gomensoro et al., 2017; Stern et al., 2010). However, sources of information available on practical exams are opinions, not based on research. Discussion addresses the relevance and fairness of the exams, which depend on factors such as apprentice supervision and the fluctuating quality of training (Kuster, 2011), and the timing of exams (Tellenbach, 2022). These studies focus on the development of vocational education

and training and the interactions between companies, schools, and vocational centers. Our research aims to address this gap by using a scientific approach to survey examiners on the validity of the practical exam.

The practical exam is a complex experience for the examiner, as it prompts them to reflect on their own professional identity. According to Figari (1994), an examiner applies referentialization during a practical exam, comparing the gap between what they observe versus would like to observe in the apprentice. This referentialization is even more complex when there is a second examiner which means results have to be negotiated, as well as an exam protocol representing the official referent, which may differ from the examiner's referent. We analyze these tensions using the model of identity transactions (Perez-Roux, 2016) and the various perceptions of validity identified by Penta et al. (2005).

Issues and objectives

A study by Kägi (2010), mandated by the Swiss Confederation, surveyed umbrella training associations, host companies, schools, and trainees. The findings showed that 25% of examiner training was considered adequate, while 18% was deemed inadequate, and that there is insufficient continuing training for examiners with intervals between successive training too long. Training courses are typically associated with reforms of legislation or exam procedures usually implemented every five to ten years. The study also raised the issue of a lack in the number of trained professionals, due to a limited population with appropriate training. Another study, using the Angoff method (Angoff, 1984; Bain, 2017) observed the degree of agreement between examiners (Jan, 2023). Examiners were asked to imagine a panel of 100 apprentices with minimal competencies, i.e. able to solve the questions in the practical exam, but with the lowest score threshold for a pass. The examiners were asked to analyze each question one-by-one and estimate how many of the 100 participants would answer it correctly. Six examiners with over 10 years of experience, who had assessed over 100 apprentices and prepared apprentices for practical exams every year, were asked to examine each practical exam question using this method. A freely accessible version of the exam protocol was used, giving both apprentices and instructors the opportunity to prepare for the practical exam. The results showed significant differences in the examiners' estimates. For example, one examiner estimated that 30 minimally competent apprentices

would answer one question correctly, while another estimated that 95 for a difference of 65 points. The average divergence for the questions was 45 points (Jan, 2023), giving an inter-examiner agreement of 0.39 according to Krippendorff's α test (Hayes & Krippendorff, 2007). These examples illustrate the high subjectivity of examiners' judgments (Kägi, 2010), even among experienced professionals using the same protocol (Jan, 2023).

The Function of the Expert

The examiners are professionals working in the food retail sector, in most cases as supermarket managers. They are required to have at least three years of practical experience and hold a diploma identical to or higher than that being assessed. To qualify as an examiner for the practical exam, professionals must attend a two-day training course in accordance with SERI's Ordinance on Vocational Training (2004). The basic course, on the first day, is common to all vocational education and training professions. It covers the examiners' legal responsibilities and duties. The second day, the specific course is delivered to examiners grouped by professional sector. It features practical information regarding the exam, including the use of exam protocols, time management, and good practices in dealing with certain situations involving apprentices.

Examiners are chosen for their examiner's attitude and field expertise in the trade (Delahais et al., 2021). As a result, they are recognized by their peers as competent professionals with a positive attitude to the profession, the qualification procedure, and apprentices.

The practical exam process

Practical exams take place in May each year, and vocational school exams are in June. This practical assessment (Brown et al., 1989; Mottier Lopez & Allal, 2008) takes place in the company where the apprentice is training.

There are two qualifications: retail assistant which requires two years of training and a 60-minute practical exam; and retail manager which requires three years of training and a 90-minute practical exam. The assistant training program is aimed at individuals with academic difficulties, and they spend less time in the classroom than apprentices in the manager training program. According to Bloom's taxonomy (1969), the questions in the assistants' practical exam are less complex and the exam is 30 minutes shorter. Apart from these two differences, the two training courses are examined by examiners in the same way and according to similar protocols.

Immediately after the exam, for about 30 minutes, the examiners negotiate and agree on the number of points to allocate to each question (Reuter et al., 2013). The final total is used to calculate the practical exam mark, which is then sent to the coordinator.

This is not the same assessment practice as for teachers who have worked with and trained with their students for several years. It is more similar to an external assessment (Merle, 2018) or an institutional assessment (Perrenoud, 2010), where professionals from the field meet to carry out their examiner function. The examiners do not know the apprentices nor the company which can lead to novel and destabilizing situations for the apprentices. For example, an examiner could ask how products are selected for the shop's promotional areas, yet the company did not include employees in the decision-making process.

The examiners assess whether the operational skills set out in the training plan have been achieved (FCS, 2021a, 2021b). These operational skills, better described by the original German term *Handlungskompetenz*, are the "skills" of "carrying out actions". In this case, rather than describing how they would carry out a task, apprentices must demonstrate their ability through an action in a situation that allows for situated assessment (Vial, 2012). These skills result from regular company-based practice and an ongoing dialogue between theory and practice (Perrenoud, 2000). The exam puts apprentices in situations typical of their day-to-day work. For instance, assessors go to the till with a selection of items and evaluate how well the apprentice handles the goods, complies with legal and safety instructions, and interacts with others while working. The level of complexity of each question is defined according to Bloom's original taxonomy (1969).

The practical exam focuses on the operational skills learnt that are required in professional practice (Le Boterf, 2016), but the assessment is not formative and is not intended to help apprentices improve (Scriven, 1996). No feedback on their performance is provided at the end of the practical exam. It is an assessment for certification to confirm that the apprentices are qualified for employment in their field.

The practical exam protocol

The basic structure of the practical exam protocol, which defines the number of questions per area, the taxonomic level, and the time allowed for each question, is produced by the Formation du Commerce de Détail Suisse. The practical exam questions are generated automatically from

a stratified sample of questions (Laveault & Grégoire, 2014) on an IT platform and can be adapted to suit different types of business. This automated and structured construction of the practical exam means that: (a) the protocol differs from one apprentice to another; (b) the protocol is adaptable to the type of business visited, for example, a small village shop, large shop, or specialist shop; and (c) the exam covers all the skills of the trade without focusing on one area in particular.

Each question (see example in Figure 1) offers a series of probable solutions in the column headed Example of solutions, as well as a free Protocol zone for noting the apprentice's answer. Examiners are free to use either or both zones. The most important requirement is recording information that retraces the progress of the practical exam when assessed by two examiners.

The other zones in Figure 1 provide administrative information, such as the source used to create the question, the time allocated, and the number of points awarded.

Theoretical Framework

As shown, the practical exam protocol provides examiners with the questions asked and a clear framework for their task. However, the practical exam process also means that examiners examine apprentices and companies they know almost nothing about. To fully grasp this situation, we assess the examiners' perceptions of the exam's validity (Penta et al., 2005). The types of validity are integrated into the model of identity transactions (Perez-Roux, 2016), showing how referentialization (Figari, 1994) by examiners in a framed context that causes instability is linked to their professional identity.

Validity of the exam

In the first phase, we test the validity the examiners attribute to the exam (Anastasi & Urbina, 1996; Laveault & Grégoire, 2014; Nevo, 1985; Penta et al., 2005). In the second phase, latent class analysis is used to produce examiner profiles based on the validity factors constructed. These profiles are then applied in the third phase to learn more about the examiners' referentialization process (Figari, 1994).

Figure 1
Question from a practical exam protocol (translated from French into English)

No. Obj.	Evaluation objective	TA	No.	Task	Min.	Pt.	Example of solutions	Protocol		Points
5.3.14	I am helpful and committed to my work. **Instructions for the examiner:** No role-playing; the examiner presents the case. **Theme:** Types of specific clients	C3	2.4	**Sales interview/case presentation** A blind customer enters your shop with their guide dog. How do you behave? What do you do in such a situation? Show us how you would act as if you were in a real-life situation. 1,0 P = very good 0,5 P = satisfactory	3	1	– Approach the customer – Introduce yourself by name as a member of staff – Ask if the customer needs assistance – Be careful with your choice of words (colours, appearance, avoid expressions such as 'do you see') – Do not hinder the dog in its work	1		

We use three out of the six types of validity (Penta et al., 2005):

1) *Content validity*: Examiners are asked about the practical exam as a quality tool for assessing the professionalism of apprentices (Laveault & Grégoire, 2014; Penta et al., 2005). For example, do the questions accurately reflect apprentices' day-to-day work?
2) *Face validity* questions the apparent quality of the exam. The examiners are retail professionals, but they are not experts in education or aware of the training plan that inspires the practical exam questions. Therefore, this validity relates to the apparent rather than the actual value of the practical exam (Penta et al., 2005). This measure is important because it indicates the legitimacy and credibility the examiners give to the practical exam protocol (Anastasi & Urbina, 1996). For example, whether a turn of phrase determines how a question is understood.
3) *Consequential validity* represents the perception of the fit between the practical exam questions and the shop where the appraisal takes place (Laveault & Grégoire, 2014; Nevo, 1985; Penta et al., 2005). This validity makes it possible to take into account the consequences of the practical exam protocol on the evaluation of the practical exam itself (Messick, 1988). For example, do the trainees' answers correspond to the list of solutions?

It should be noted that these three validities are questioned according to the examiners' perceptions of the validity of the practical exam. It is not the exam itself being tested, but rather the approach to the practical exam to correctly perform the examiner's function (Loye, 2019).

The other three types of validity (Penta et al., 2005) refer to concomitant, predictive, and conceptual criteria and do not involve questioning the examiners about the exam. These criteria question the validity of the practical exam compared with other exams, and as a result, are not relevant to this study.

Examiner profiles

The procedure for constructing profiles using latent class analysis is presented in the results section. Going forward, we refer to these different profiles using Figari's (1994) approach.

During the assessment, the examiner relies on a frame of reference (Figari, 1994, 2006), which is not a pre-existing reference object but an ephemeral, personal construct that is cognitively mobilized during the exam. This frame of reference is conceived as the standard tool used to assess the gap between the referent, which designates what we would like to observe or note, and the referred, which corresponds to the reality observed during the practical exam in the field. The referent is multidimensional, incorporating the examiner's expectations according to their personal quality criteria, market needs, the type of company visited, and their state of mind at the time. Therefore, the very conception of the examiner function (Issaieva & Crahay, 2010) is specific to each individual. Examiners examine apprentices based on what they show and demonstrate during the practical exam, defined by Hadji (2012) as "talking signs". These signs or clues are derived from the professional practices (Le Boterf, 2016) the examiners want to see in the apprentice's performance.

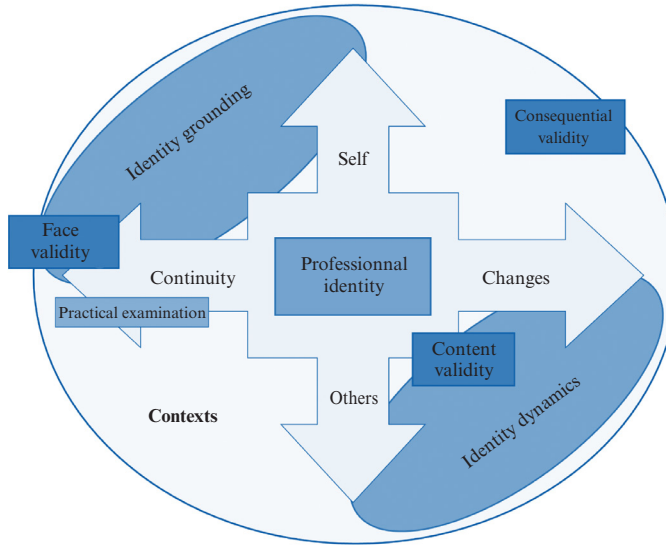
The practical exam has a certification function, but the approach to this assessment can be either summative or hermeneutic (De Ketele, 2010): Summative because obtaining a sufficient mark is compulsory in order to obtain a diploma, but also hermeneutic because the practical exam makes sense in terms of professional practices to guarantee the skills of a future qualified employee (Mottier Lopez, 2015). There is also bias due to unfamiliarity with the company where the assessment takes place, as examiners are never familiar with the apprentice's company. Thus, a "company effect" analogous to the "school effect" (Grisay, 2006) may occur, whereby examiners are influenced by a company's reputation. Hence, an examiner's judgement (Bressoux, 2018) can be influenced by prejudices about a company which colors the assessment of the apprentice.

Therefore, referencing (Figari, 1994) is the most important part of the assessment process because the result is not justified in absolute terms. On the contrary, it is relative to the context of a specific practical exam according to the difference measured between the referent and the referred.

Professional identity

The referent (Figari, 1994) can be twofold: On the one hand, the examiner, and on the other, the official referent which is the practical exam. To explain how these referents coexist, we use an adapted version of Perez-Roux's (2016) model of identity transactions (Figure 2), which was originally used to present the complex and dynamic process of coaching in training.

Figure 2
Identity transactions (adapted from Perez-Roux, 2016, p. 4)



During the assessment, the examiner's professional identity is divided between several dimensions. Precision "during the exam" is important because the examiner is not in their usual context, in an unfamiliar company to examine an unknown apprentice with a second examiner, also often known only due to this process. Consequently, the examiner's professional identity is not fixed during the assessment. The examiner's referentialization (Figari, 1994) fluctuates according to the dimensions and contexts of identity transactions (Perez-Roux, 2016). According to the horizontal axis, to guarantee the conformity of the official assessment, the examiner must refer to the fixed structure of the practical exam while, at the same time, assessing whether the apprentice's skills correspond to the current criteria of the trade. The vertical axis calls for adjustments to be made between the apprentice's own image of what the trade should be, the examiner's criteria for a job well done, and what is required by the exam, as well as the skill expectations of the examiner colleague for the trade. Referentialization (Figari, 1994) is the association the examiner makes between their various identity transactions, their examiner colleague, the exam protocol, and the apprentice being examined.

In the model (Figure 2), the practical exam is an invariable object on the "Continuity" axis. We also position the three types of validity retained. Face validity, in the Practical exam zone, specifically assesses the construction of exam protocols according to the questions and turns of phrase included. Second, according to the fit between the practical exam protocol and the assessment of apprentices at the end of their training, Content validity is located between the Changes and Others axes, as it also evaluates the use and application of the protocol by the second examiner. Third, Consequential validity evaluates the tool's consistency between the questions, the type of store, and the training program, for evaluating a future professional. This validity type is located between the Changes and Self axes, as it refers to the examiner's perceived difference between their personal referent (Figari, 1994) and the official referent i.e. the exam protocol.

Research objectives and hypotheses

It should be possible to extract the three dimensions of the validity of the qualification procedure from the answers to the questionnaire items, and use them to answer our research question: How do examiners differ according to their assessment of the validity of the practical exam?

Our first hypothesis is that examiners can be categorized into different profiles according to the three dimensions of validity. We tested for these profiles using latent classes. Our second hypothesis is that, depending on their profile, examiner referentialization is largely based on their experience. The role of examiner is occasional, with little supervision, and the structure of the practical exam protocols has not been modified since 2004. As a reminder, examiner training lasts two days and takes place when they begin as examiners. Following a change in taxonomic level in 2014, examiners were offered a refresher course on the second day, although some had not taken the full course for 15 years. We will test the robustness of this hypothesis using other sociodemographic data and certain referencing items during the practical exams.

Method

Participants

The population of 457 examiners in the food sector was canvassed using the national database of examiners. Although the qualification procedure is a canton responsibility, the examiners were divided into three groups according to French-, German-, and Italian-speaking regions.[2]

The sample (n = 189) comprised 32% females, with an average age of 46 (M = 46.48, SD = 9.96). This represents 41.35% of the population. According to the International Standard Classification of Education (ISCED), 28% of respondents had completed education at Level 5 and 72% at Level 6. French speakers accounted for 36% of the sample, compared with 22% of the population. German speakers accounted for 59% of responses (compared with 75% of the population), and Italian speakers accounted for 5% (compared with 3% of the population). The examiners had an average of 11 years of experience (M = 11.19; SD = 6.92).

According to an internal report on the food retail sector carried out in 2023, examiners carry out an average of 10 appraisals, taking approximately three working days at an average rate of three to four appraisals per day.

Responses received in under six minutes were removed from the sample (n = 6), as the minimum time required for a quality response according to empirical tests with three participants.

Instruments

Eleven Likert-type items were constructed to assess perceptions of content validity, face validity, and consequential validity (Penta et al., 2005). The response modalities varied across six levels, ranging from "totally disagree or does not correspond to me at all" to "totally agree or corresponds to me completely".

Part 1: General questions about the practical exam

In Part 1, the examiner was requested to indicate their degree of agreement with general statements about the practical exam. For example, "Generally speaking, you feel that the exam lasts too long" and "In your opinion, the practical exams and exam protocols reflect the day-to-day work of apprentices".

2. The choice of language, French, German, or Italian, is defined in the apprentice's employment contract which determines the course and practical exam language.

Part 2: Questions on how to answer

In Part 2, Examiners were asked to indicate their degree of agreement with questions and statements about screenshots of specific points in the free version of the practical exam.

Part 3: Sociodemographic data

The requested information included the employee's gender, age, highest level of education, and the company where they worked. Companies were divided into two groups: "Leader and Historic, which refers to two companies that account for 64% of retail sales according to the Federal Statistical Office (2024), and Other, which refers to all other companies.

We asked the companies how many years their examiners had worked, the average number of examiner assessments they had conducted, and the number of apprentices in their companies. Examiners' experience was divided into three categories: less than 3 years (little or no experience), 3 to 15 years (some experience but who had only applied the 2004 practical exam protocols), and more than 15 years (substantial experience who have worked with both 2004 practical exam protocols and the previous ordinance). It is important to distinguish between the last two groups, as the previous professional training ordinance imposed a duration but not an exam protocol. Under the previous ordinance issued by the Federal Office for Professional Education and Technology (OFFT, 1980), examiners were free to ask apprentices questions of their choice.

The instrument was translated into Italian and German. The Italian version was validated by the head of the Italian region. The German translation was validated using the back-translation method (Behling & Law, 2000), and only remarks concerning abbreviations of titles and institutions were noted. The canton in question was not pertinent, as examiners are often required to assess several cantons.

The three types of exam validity

We used four content validity items (Laveault & Grégoire, 2014; Penta et al., 2005) to measure acceptance of the practical exam protocol as an appropriate assessment tool for apprentices. The protocol determines the entire exam procedure and questions asked by examiners. Therefore, the protocol must reflect daily work practices and be a practical tool.

Face validity, measured by three items, is considered less rigorous (Anastasi & Urbina, 1996; Nevo, 1985; Penta et al., 2005) but is important as it measures the degree to which the examiners accept and trust the test. Although all the questions in the protocol are constructed using their designers' terminology, the professional language, technical terms, and turns of phrase varied from one training company to another. For instance, depending on the company, the same department may be referred to as "grocery", "general merchandise", or "dry goods". Therefore, this validity type measured the credibility that the examiners attributed to the terms chosen but also their assessment of the apprentice's ability to understand the different terms.

Finally, four items were used to measure the implications attributed by the examiners to the protocol as a measurement instrument: consequential validity (Laveault & Grégoire, 2014; Messick, 1988; Penta et al., 2005). The score the examiner assigned should indicate whether they felt that the apprentice was competent enough to practice as a qualified professional.

Procedure for administering the instrument

The instrument was sent to the examiners by email in the language determined by the examiner's selection in the sector's IT system (French, German, or Italian). The instrument, which was available in three languages, was supplemented by the LimeSurvey online survey software. During the process, the examiner could select the interface language. The examiner was matched to one of the three languages according to the language used to complete the survey.

Examiners were requested to answer all the questions. Instructions on the survey's objectives were provided in the email and in LimeSurvey. The tool was accessible for one month, and a reminder email was sent after two weeks.

Analysis procedure

We conducted exploratory factor analysis on the 11 validity items. The three factors extracted were analyzed using Mplus 8.3 software to produce latent classes and Kruskal–Wallis tests were used to compare the compositions of the resulting three examiner profiles.

Results

Our research had several objectives. First, according to our theoretical framework, we aimed to extract the three validity dimensions of the qualification procedure. Second, we aimed to distinguish examiner profiles according to these validity dimensions. Finally, we aimed to interpret these profiles using the sociodemographic data collected.

Exam validity

Exploratory factor analyses were conducted using principal axis factorization extraction and oblimin rotation with Kaiser normalization (Stafford et al., 2006). Based on a sample of 189 participants, only factor loadings greater than 0.37 were considered (Berger, 2021). The three extracted factors explained a total of 57.59% of the variance. Bartlett's sphericity test was significant, and the KMO index was adequate at 0.70 (Gagnon, 2019). Following the advice of Béland et al. (2018), we report internal consistency, using McDonald's omega, as the perception of the validity of practical exams is multidimensional depending on the context of the exam as well as the referent constructed by the examiner.

The first factor extracted corresponded to content validity, comprised four items, had an eigenvalue of 2.85, and explained 25.89% of the variance in the items. Internal consistency was good, with a McDonald's omega value of 0.73. The item "I can easily award 0 to 17 marks for presentation" was retained, despite being only slightly saturated, as the 17 possible marks for this question carry significant weight in the practical exam for two reasons. First, 15 minutes are allocated to this question, representing one-sixth of the total time. Second, this question is paired with four others that can give up to eight extra points. Therefore, this question is worth 25 points, a quarter of the total marks available in the exam.

The second extracted factor corresponded to face validity, comprised three items, had an eigenvalue of 2.21, and explained 20.07% of the variance in the items. Internal consistency was good, with a McDonald's omega of 0.79.

The third factor extracted corresponded to consequential validity, comprised four items, had an eigenvalue of 1.28, and explained 11.63% of the variance in the items. Internal consistency was relatively weak, with a McDonald's omega of 0.54.

Examiner profiles

The three validity factors were standardized into Z scores and processed in Mplus 8.3 software (Asparouhov & Muthen, 2012; Caron, 2018; Geiser, 2013; Nylund et al., 2007) to extract examiner profiles.

Table 1
Pattern matrix of the three-factor solution

Dimensions	Item	Factor		
		F1	F2	F3
Content validity ω = 0.73	Sufficient to certify that the apprentice is a competent professional.	0.89		
	Reflects the day-to-day work of apprentices.	0.63		
	Has a level of difficulty adapted to certify the skills of a retail manager.	0.63		
	I can easily award 0 to 17 points for presentation.	0.37		
Face validity ω = 0.79	The technical terms used make the questions difficult.		0.78	
	The language used makes the questions difficult.		0.75	
	The phrasing makes the questions difficult.		0.72	
Consequential validity ω = 0.54	Apprentices' answers correspond to the list of solutions.			0.63
	I can easily distinguish between the three skills (professional, methodological, and social) to be examined when analyzing the presentation.			0.51
	The answers in the list of solutions seem to correspond to current retail practices.			0.47
	The apprentices' presentations seem to be generally consistent with the theory.			0.38
Eigenvalue		2.85	2.21	1.28
Variance explained (%)		25.89	20.07	11.63

Note: Loadings < 0.37 are not displayed

Table 2
Statistical indices for choosing the number of profiles to retain

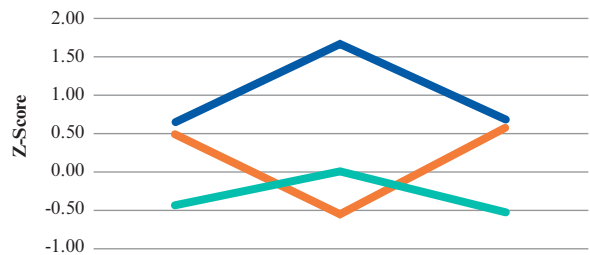
No. of classes	Log(L)	AIC	BIC	ABIC	BLRT	Entropy
2	-803.03	1 601.31	1 633.72	1 602.05	< 0.001	0.46
3	**-790.65**	**1 587.84**	**1 633.22**	**1 588.88**	**< 0.001**	**0.60**
4	-779.92	1 583.66	1 642.01	1 585.00	ns	0.69

Log(L), log likelihood; AIC, Akaike Information Criterion; BIC, Bayesian Information Criterion; ABIC, Adjusted Bayesian Information Criterion; BLRT, Bootstrap Likelihood Ratio Test

Following the recommendations of Nylund et al. (2007) and using the Bootstrap Likelihood Ratio Test (BLRT), recognized for its robustness in this context, we retained three latent classes, shown in Table 2). Nylund et al. (2007) suggested retaining the final latent class, which produced a significant result ($p < 0.01$). The value of the Bayesian Information Criterion (BIC) test reinforced this decision, as it was lower for the three-class model, highlighting the relevance of this configuration.

Figure 3 shows the averages for each profile type by validity type.

Figure 3
Averages for validity type according to examiner profiles



	Content validity ηp2 = 0.37	Face validity ηp2 = 0.30	Consequential validity ηp2 = 0.46
1. Low-validity practical examination (n = 105)	-0.47	-0.01	-0.56
2. Low-legitimacy practical examination (n = 63)	0.48	-0.56	0.60
3. High-validity practical examination (n = 21)	0.63	1.65	0.66

Table 3

Description of examiners belonging to each profile

	n (%)	Level of education		Company[1]		Examiner experience			Gender		Language			Age
		ISCED 5	ISCED 6	Leader and Historic	Other	Less than 3 years	3 to 15 years	More than 15 years	F	H	French	German	Italian	
		(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	(%)	
1. Low-validity practical exam	105	28	77	69	36	10	72	23	39	66	29	70	6	46
	(55.56%)	(53.85%)	(56.20%)	(58.47%)	(50.70%)	(55.56%)	(55.38%)	(56.10%)	(63.93%)	(51.56%)	(42.65%)	(62.50%)	(66.67%)	
2. Low-legitimacy practical exam	63	15	48	39	24	4	43	16	16	47	23	37	3	48
	(33.33%)	(28.85%)	(35.04%)	(33.05%)	(33.80%)	(22.22%)	(33.08%)	(39.02%)	(26.23%)	(36.72%)	(33.82%)	(33.04%)	(33.33%)	
3. High-validity practical exam	21	9	12	10	11	4	15	2	6	15	16	5	0	46
	(11.11%)	(17.31%)	(8.76%)	(8.47%)	(15.49%)	(22.22%)	(11.54%)	(4.88%)	(9.84%)	(11.72%)	(23.53%)	(4.46%)	(0.00%)	
Total	189	52	137	118	71	18	130	41	61	128	68	112	9	

Note: Percentages add up to 100% when read as a column.

1. The companies are divided into two groups: Leader and Historic, which, as the name suggests, correspond to the two historic and leading companies in Switzerland, and Other for all the other companies.

Table 3 shows the breakdown of each profile by level of education, company, examiner experience, gender, language, and age. The profile composition differs in language and experience as an examiner. The distribution of examiners in the profiles depends on their level of experience ($F_{(1,186)} = 3.15$; $p < 0.05$; $\eta 2 = 0.03$) and language ($\chi^2_{(2)} = 16.22$; $p < 0.001$; $V = 0.30$). Due to the small number of Italian speakers (n = 9), the test was conducted with German and French speakers instead. However, we found that this distribution was independent of the examiner's level of education, company, gender, and age.

Description of Profile 1: Low-validity practical exam

This profile covers 56% of the examiners in the sample: 105 out of 189 examiners. They rated the practical exam procedure as having low validity in terms of both content validity and consequential validity. Regarding face validity, the examiners' Z-score (-0.01) was almost exactly average for the total sample. Face validity attributed to the practical exam (Anastasi & Urbina, 1996; Nevo, 1985; Penta et al., 2005) was within the average for the other examiners. However, both content validity (Laveault & Grégoire, 2014; Penta et al., 2005), for the practical exam's capacity to evaluate professionals, and consequential validity (Laveault & Grégoire, 2014; Penta et al., 2005) for the validation process put in place by the exam, were lower than for the two other profiles.

Regarding language and experience, this profile was adopted by more German speakers (63%) than French speakers (43%). Average experience is 11 years (M = 11.45; SD = 7.11), which is similar to that for the whole sample.

Description of Profile 2: Low-legitimacy practical exam

This profile covers 33% of the examiners in the sample: 63 examiners. They gave a positive assessment of content and consequential validity (Penta et al., 2005) and considered the practical exam to be satisfactory in substance. However, face validity was below average, and the examiners gave little credence to the construction and wording of the practical exam (Anastasi & Urbina, 1996; Nevo, 1985; Penta et al., 2005). While the protocol was perceived as an adequate tool for carrying out the exam, the terms used in it were not considered appropriate.

This profile is composed of an equal number of German (33%) and French (34%) speakers. The average experience of this profile is 11.92 years (SD = 6.54), similar to the general average.

Description of Profile 3: High-validity practical exam

This profile covers 11% of the examiners in the sample: 21 examiners. According to the three constructed indicators (Anastasi & Urbina, 1996; Laveault & Grégoire, 2014; Messick, 1988; Penta et al., 2005), the practical exam was perceived as highly valid. The scores were the highest of the three profiles. The content validity and consequential validity scores were similar to those of Profile 2 and higher than the Profile 1 scores. Face validity scores were notably higher than those of the other two profiles, with a Z-score of 1.65. The examiners in this profile attributed a high level of validity to the exam.

This profile consists almost exclusively of French speakers (24%), with only 4% German speakers, and no Italian speakers. Experience is lowest, at eight years (M = 7.71; SD = 9.29).

Our first hypothesis is verified: Three examiner profiles, based on how the exam is validated, can be identified.

Sociodemographic characteristics of examiner profiles

In line with our second hypothesis, we analyzed the examiner profiles using different sociodemographic data to determine whether experience as an examiner is the primary source of referentialization (Figari, 1994, 2006).

Examiner profile and experience

As the data did not follow a normal distribution, we used the non-parametric Kruskal–Wallis test. This test shows whether there are differences between the examiners' profiles and the observed data. The mean rank indicates the observed group's position in relation to the expected mean rank.

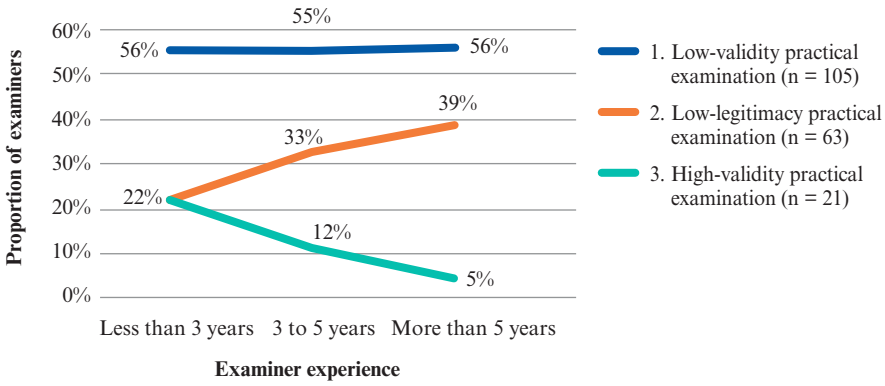
Table 4
Examiner profiles by average experience

Group	n	M	SD	Average rank
3. High-validity practical exam	21	7.71	6.29	65
1. Low-validity practical exam	105	11.45	7.11	97
2. Low-legitimacy practical exam	63	11.92	6.54	102

Table 4 shows a significant difference in evaluations of the legitimacy of the exam ($H(2) = 7.49$; $p < 0.05$) (Siegel, 1956) for an effect size of 4% between the examiners' profiles and experience. The difference between the expected mean rank (95) and the three mean ranks shows that examiner experience did not have an identical central tendency across profiles. Therefore, examiner experience, as defined by the constructed profiles, effectively represents distinct populations.

Examiners in Profile 1 (Practical exam with low validity) and Profile 2 (Low-legitimacy practical exam), indicating a perspective that partly challenges the validity of the exam, had longer experience in this role. Figure 4 shows the three examiner profiles according to experience groups: Less than 3 years, 3 to 15 years, and More than 15 years. Profile 1 is similar across the three experience groups. Profiles 2 and 3 are identical at 22% in the group of examiners with less than three years' experience but diverge in the other two groups. For examiners with more than 15 years of experience, Profile 3 accounted for only 5% of examiners, whereas 39% were in Profile 2. It should also be noted that our data are cross-sectional and not longitudinal, so it is possible to make comparisons between examiner profiles but not observe changes in them.

Figure 4
Examiner profile by experience



This information is supported by four other results. First, when the Kruskal–Wallis test was performed with age instead of experience, the result was insignificant ($H(2) = 0.81$; *ns*), indicating that years of experience and not age are linked to the examiners' profiles. Secondly, a Kruskal–Wallis test comparing the three examiner profiles ($H(2) = 6.80$; $p < 0.05$) that excluded the item "The length of the exam [...] is too long", was also significant. This means that the test was perceived as too long by the "High-validity practical exam" (rank mean of 103) and "Low-validity practical exam" (rank mean of 102) classes but not by the "Low-legitimacy practical exam" class (rank mean of 81). For examiners with more experience, the duration of the practical exam was considered too short to make a valid assessment. Two chi-square tests revealed that the profiles did not differ in terms of educational composition ($\chi^2_{(2)} = 2.29$; *ns*) or ($\chi^2_{(2)} = 2.63$; *ns*).

Examiner profile by language

There was a significant difference according to the examiners' languages ($\chi^2_{(2)} = 16.22$; $p < 0.01$; $V = 0.30$). The chi-squared test was performed on French- and German-speaking examiners. Italian speakers could not be included, as the theoretical distribution table showed a value of less than 5. Table 5 shows the distribution of the three profiles according to language.

Table 5
Examiner profile by language

Group	n	French	German
1. Low-validity practical exam	105	43%	63%
2. Low-legitimacy practical exam	63	34%	33%
3. High-validity practical exam	21	23%	4%

Profile 2 had identical proportions of the two languages, while Profile 3 seems to be present in the majority of French-speaking examiners.

Discussion

We used exploratory factor analysis to extract three dimensions of the validity of the qualification procedure. These dimensions then enabled us to distinguish three examiner profiles using latent classes analysis. Finally, we examined the relationship between these profiles and the sociodemographic data collected. This discussion focuses on the significance of the

examiner profiles and their perceptions of the practical exam's validity, and how these perceptions are modulated by the examiners' experience and language.

Validity of the practical exam according to examiner profiles

According to the three types of validity tested, the majority of examiners considered that the practical exam was not very suitable. From the perspective of content validity (Laveault & Grégoire, 2014; Penta et al., 2005), the practical exam protocol was not considered an effective means of assessing apprentices' skills. According to consequential validity (Laveault & Grégoire, 2014; Nevo, 1985; Penta et al., 2005), it was not considered an adequate tool for assessing the examinees of various companies. Finally, for face validity (Anastasi & Urbina, 1996), the practical exam was not perceived as legitimate because of the turns of phrase used.

We hypothesize that this low assessment of the practical exam's validity is attributable to its lack of evolution, due to the need for continuity of identity grounding (Perez-Roux, 2016). The text of the law and the training plan have not been modified since 2004, so the practical exam has not evolved. The examiners, on the other hand, especially those who have held this position since 2004 or even earlier, have experienced significant changes in their professional identities in the retail trade and therefore in their day-to-day work, which is not reflected in the current exam protocols. We also proposed the hypothesis that examiners experience lassitude due to using identical practical exam protocols year after year. Experienced examiners are so familiar with the unchanged exam, they do not even need to refer to it anymore, possibly giving them the impression that their own feelings are more valid than those generated by a protocol that has become almost superfluous to their practice.

According to examiner experience

Based on the exam validity factors, three examiner profiles significantly related to examiner experience were constructed.

Profile 1 (Low-validity practical exam), which describes the practical exam as weakly valid, is similar across the three experience groups. It is also interesting to observe the difference between the Profiles 2 and 3, the low and high validity practical exam profiles. For examiners with less than three years of experience, the split between Profile 2 and Profile 3 is

identical, at 22%. The assessment of the practical exam as being highly valid (Profile 3) decreases in the next two experience groups, and vice versa for the assessment of the practical exam as not very valid (Profile 2).

Our data are cross-sectional and not longitudinal, pointing to a difference between different examiner experience, rather than evolution. The third profile decreases to the same degree as the second increases. The difference is largely explained by the face validity of the exam. The more years of experience, the lower the legitimacy and credibility (Anastasi & Urbina, 1996) attributed to the practical exam protocol by examiners. We put forward two arguments to explain this difference.

First, we recall important three points regarding the examiners' role: (a) they receive a single two-day training course at the start of their careers; (b) no changes have been made to the practical exam protocol since 2004; and (c) exams take place for an average of two days per year. Consequently, over time, examiners have developed their own conception of the examiner function and the referentialization used in their assessments (Issaieva & Crahay, 2010; Figari, 1994). The referent—what is to be observed—should remain the same; however, referentialization is a dynamic process constructed by the gap between the referent and the referenced, representing the reality observed by the examiners. This gap evolves over time as company practices, which were previously unknown, become familiar. Consequently, examiners develop a frame of reference where the original reference fades, and the credibility attributed to a non-evolving procedure diminishes as their experience increases.

Second, this assessment of the low validity of the exam, which was higher according to the examiners' experience, correlated significantly with the item describing the practical exam as too short. This correlation indicates a discrepancy in the examiners' assessments. We are moving away from a strictly summative approach, in which a mark is awarded for a practical exam, toward a hermeneutic approach (De Ketele, 2010), in which the examiner assesses the quality of reflection in action (Perrenoud, 2001). The identification of meaningful indicators (Hadji, 2012) in professional practices (Le Boterf, 2016), considered alongside the examiners' growing experience, is a time-consuming process.

Other indicators support this analysis based on the examiners' experience. As a reminder, gender, age, level of education, and type of company were not found to be significantly linked to examiner profiles. Only language (French or German) showed a significant difference. This will be discussed in the following section.

According to examiner language

As shown in Table 5, around a third of examiners in both languages assessed the practical exam as having little legitimacy. The main difference is in the profile describing the practical exam as highly valid, which represents 4% of German-speaking examiners but 23% of French-speaking examiners.

We attribute this to the way examiners begin their functions in the different regions. French-speaking examiners start out as examiners as soon as they are appointed, whereas German-speaking examiners start out as observers of other examiners, who serve as models for them during their first year. For instance, a French-speaking examiner who has just finished their training conducts an exam alongside a second examiner and plays an active role in the evaluation process. In contrast, for German-speaking examiners, the first year of observing a pair of model examiners gives novice examiners the opportunity to observe without taking responsibility for the assessment.

This Alemannic method, involving a year of observation, allows more exchanges on formal and informal practices for a good practical exam between the two experienced examiners and the novice examiner. However, model examiners are chosen from among the most experienced, who, as we saw earlier, rate the practical exam as having low validity. Consequently, novice examiners may be influenced by the more experienced examiners' opinions about the exam.

Limitations and avenues for future study

Sampling the population shows that French speakers are overrepresented compared to German speakers, which gives them greater weight in the representation of the data. The conclusions may differ with a more representative sample based on the distribution of languages at the federal level. However, this difference is not significant in the results obtained, and we only find a more marked presence of French-speaking examiners in the profile which perceives the practical exam as highly valid.

Second, although food training accounts for 28.56% of exams in the retail sector (Federal Statistical Office, 2023), our research is specific to one trade. A larger-scale study would distinguish between results specific to this context and those general to the sector.

Another limitation is the projection of changes in examiner profiles according to experience. As our study is cross-sectional and not longitudinal, we cannot determine how examiners' assessments of exam validity are changing. Nevertheless, we observe that the current panel of examiners has significant tendencies to assess the validity of exams depending on their level of experience. The study would need to be repeated several times to confirm or refute this.

The three constructed profiles, together with the examiners' levels of experience, will be used to focus attention and create groups for two focus group sessions. This study will provide an opportunity to focus discussions on the examiners' conceptions of their roles while simultaneously revealing the social representations of their functions and practices through interaction. The focus groups will also provide an opportunity for the examiners to discuss their recommendations for indicators of apprentice competencies.

Conclusion

Our study showed that examiners assess the validity of the practical exam as poor. Taking experience as a benchmark, our study shows that examiners with at least 15 years of experience almost never perceive the practical exam as highly valid. We hypothesized two factors: (a) lack of ongoing training for examiners; and (b) a mismatch between a practical exam unchanged since 2004 and the constantly changing identity dynamics of professionals.

Ideally, exam protocols should be adapted regularly, and ongoing training provided.

This solution is complex due to both external and internal factors. Externally, due to the scale of change and its impact. The exam is inextricably linked to established vocational education and training programs and represents its culmination. Changing the exam would require adaptation of the preparation process, including training programs in inter-company courses, vocational schools, and host companies. Internally, in the event of

changes, examiners would require ongoing training for conducting exams. The examiner function is an occasional occupation, with some examiners only working for one or two days one year and not necessarily for consecutive years. Consequently, it would be difficult to always have the required population of examiners for practical exams, as they would need to be both available and have received updated training for the current exam.

Proofreading: Caroline Lefour

Formatting: Emmanuel Gagnon

Portuguese abstract: Eusebio Andre Machado

LITERATURE CITED

Anastasi, A., & Urbina, S. (1996). *Psychological testing* (7th edition). Pearson.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Educational Testing Service.

Asparouhov, T., & Muthen, B. (2012). Using Mplus TECH11 and TECH14 to test the number of latent classes. *Mplus Web Notes*, *14*(22), 1-17.

Bain, D. (2017). Fixer un seuil de suffisance pour une épreuve de maîtrise: Apports et limites de la méthode d'Angoff. *Évaluer. Journal international de recherche en éducation et formation*, *3*(3), 69-95.

Behling, O., & Law, K. S. (2000). *Translating questionnaires and other research instruments*. SAGE Publications, Inc. https://doi.org/10.4135/9781412986373

Béland, S., Cousineau, D., & Loye, N. (2018). Utiliser le coefficient omega de McDonald à la place de l'alpha de Cronbach. *McGill Journal of Education*, *52*(3), 791-804. https://doi.org/10.7202/1050915ar

Berger, J.-L. (2021). *Analyse factorielle exploratoire et analyse en composantes principales: Guide pratique*. https://doi.org/10.13140/RG.2.2.16206.18246

Berger, J.-L., Wenger, M., & Sauli, F. (2021). What constitutes quality in the Swiss initial vocational education and training dual system: An apprentice perspective. In M. J. Chisvert Tarazona, M. Moso Diez, & F. Marhuenda Fluixá, *Apprenticeship in dual and non-dual systems: Between tradition and innovations*. (pp. 79-103). Peter Lang.

Berger, J.-L., Wenger, M., & Sauli, F. (2023). Engagement en formation professionnelle initiale duale et perceptions de la qualité de la formation. *Didactique*, *4*(2), 33-64. https://doi.org/10.37571/2023.0203

Bloom, B. S. (1969). *Taxonomie des objectifs pédagogiques* (M. Lavallée, Trad.). Education Nouvelle.

Bressoux, P. (2018). Comment se fabrique le jugement des enseignants? *Regards croisés sur l'économie*, *22*(1), 15-23. https://doi.org/10.3917/rce.022.0015

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. Educational Researcher, 18(1), 32-42. https://doi.org/10.3102/0013189X018001032

Caron, P.-O. (2018). La modélisation par équations structurelles avec Mplus. Presses de l'Université du Québec.

De Ketele, J.-M. (2010). Ne pas se tromper d'évaluation. Revue française de linguistique appliquée, 15(1), 25-37. https://doi.org/10.3917/rfla.151.0025

Delahais, T., Devaux-Spatarakis, A., Revillard, A., & Ridde, V. (2021). Introduction : Qui évalue et comment ? In *Évaluation. Fondements, controverses, perspectives*. Éditions science et bien commun. https://doi.org/10.5281/zenodo.6336071

FCS. (2021a, mai 18). *Plan de formation Assistant / Assistante du commerce de détail avec attestation fédérale de formation professionnelle (AFP)*. https://www.bds-fcs.ch/fr/Medias-digitaux/Download-Center?category=27&disposition=attachment&download=16&searchTerm=taxonomie

FCS. (2021b, mai 18). *Plan de formation Gestionnaire du commerce de détail avec certificat fédéral de capacité (CFC)*. https://www.bds-fcs.ch/fr/Medias-digitaux/Download-Center?category=27&disposition=attachment&download=16&searchTerm=taxonomie

Federal Statistical Office. (2023). *Formation professionnelle initiale: Tableaux de base - 2022 | Tableau* (No. su-f-15.03.02.01.01_2022). https://www.bfs.admin.ch/asset/fr/24468965

Federal Statistical Office. (2024). *Statistique des chiffres d'affaires du commerce de détail—Séries annuelles* (No. px-x-0603020000_104). Federal Statistical Office. https://www.bfs.admin.ch/bfs/fr/home.assetdetail.30546933.html

Federal Statistical Office, Gestionnaire de vente, 70200 (1980).

Federal Statistical Office. (2023). *Résiliation du contrat d'apprentissage, réentrée, statut de certification—Résultats pour la formation professionnelle initiale duale (AFP et CFC), édition 2023* (Nos. 1642-2300). Federal Statistical Office. https://www.bfs.admin.ch/asset/fr/29645528

Figari, G. (1994). *Evaluer : Quel référentiel?* De Boeck.

Figari, G. (2006). Les référentiels entre théorie et méthodologie. In G. Figari & L. Mottier Lopez (Éds.), *Recherche sur l'évaluation en éducation : Problématiques, méthodologies et épistémologie : 20 ans de travaux autour de l'ADMEE-Europe* (pp. 101-108). L'Harmattan.

Frey, M. (Éd.). (2010). Il y a encore des lieux de formation qui ne collaborent pas bien : Évaluation des procédures de qualification dans la formation professionnelle initiale. *Folio : la revue BCH | FPS pour les enseignants de la formation professionnelle*, *5*, 4-7.

Gagnon, R. J. (2019). Measurement theory and applications for the social sciences. *Measurement: Interdisciplinary Research and Perspectives*, *17*(4), 209-210. https://doi.org/10.1080/15366367.2019.1610343

Geiser, C. (2013). *Data analysis with Mplus*. The Guilford Press.

Gomensoro, A., Meyer, T., Hupka-Brunner, S., Jann, B., Müller, B., Oesch, D., Rudin, M., & Scharenberg, K. (2017). *Situation professionnelle à l'âge de trente ans. Mise à jour des résultats de l'étude longitudinale TREE*. TREE.

Grisay, A. (2006). Réflexions sur l'« effet-école ». In G. Figari & L. Mottier Lopez (Éds.), *Recherche sur l'évaluation en éducation : Problématiques, méthodologies et épistémologie : 20 ans de travaux autour de l'ADMEE-Europe* (pp. 34-43). L'Harmattan.

Hadji, C. (2012). *Faut-il avoir peur de l'évaluation?* De Boeck Supérieur.

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77-89. https://doi.org/10.1080/19312450709336664

Issaieva, É., & Crahay, M. (2010). Conceptions de l'évaluation scolaire des élèves et des enseignants: Validation d'échelles et étude de leurs relations. *Mesure et évaluation en éducation*, *33*(1), 31-61. https://doi.org/10.7202/1024925ar

Jan, D. (2023, juin 8). *Divergences inter-juges: En utilisant le protocole d'examen de fin d'apprentissage en alternance dans le commerce de détail*. Printemps de la Recherche en Éducation 2023 - Réseau des INSPÉ, Paris. https://doi.org/10.5281/ZENODO.7937207

Kägi, W. (2010). *Evaluation des procédures de qualification dans la formation professionnelle initiale*. Office fédéral de la formation professionnelle et de la technologie. https://edudoc.ch/record/38694/files/Schlussbericht_f.pdf

Kriesi, I., Neumann, J., Schweri, J., Kuhn, A., & Baumeler, C. (2016). *Rester ? S'en aller ? Recommencer ? Fréquence, causes et répercussions des résiliations de contrats d'apprentissage". Observatoire suisse de la formation professionnelle*. Institut fédéral des hautes études en formation professionnelle. https://www.hefp.swiss/sites/default/files/downloads/obs_trendbericht_lva_20160826_fr.pdf

Kuster, H. (2011). Des procédures de qualification différentes pour les apprentis. *La revue BCH | FPS pour les enseignants de la formation professionnelle*, *6*, 18.

Laveault, D., & Grégoire, J. (2014). *Introduction aux théories des tests en psychologie et en sciences de l'éducation* (3e éd). De Boeck.

Le Boterf, G. (2016). *Professionnaliser : Construire des parcours personnalisés de professionnalisation* (7e éd.). Eyrolles-Éd. d'Organisation.

Loye, N. (2019). Et si la validation était plus qu'une suite de procédures techniques? *Mesure et évaluation en éducation*, *41*(1), 97-123. https://doi.org/10.7202/1055898ar

Merle, P. (2018). *Les pratiques d'évaluation scolaire : Historique, difficultés, perspectives*. PUF.

Messick, S. (1988). Meaning and values in test validation: The science and ethics of assessment. *ETS Research Report Series, 1988* (2), i-28. https://doi.org/10.1002/j.2330-8516.1988.tb00303.x

Mottier Lopez, L. (2015). *Évaluations formative et certificative des apprentissages : Enjeux pour l'enseignement*. De Boeck.

Mottier Lopez, L., & Allal, L. (2008). Le jugement professionnel en évaluation : Un acte cognitif et une pratique sociale située. *Revue suisse des sciences de l'éducation*, *30*(3), 465-482. https://doi.org/10.24452/sjer.30.3.4798

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement*, *22*(4), 287-293. https://doi.org/10.1111/j.1745-3984.1985.tb01065.x

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535-569. https://doi.org/10.1080/10705510701575396

Penta, M., Arnould, C., & Decruynaere, C. (2005). *Développer et interpréter une échelle de mesure: Applications du modèle de Rasch*. Mardaga.

Perez-Roux, T. (2016). Accompagnement et reconnaissance d'autrui : Quels enjeux pour l'évaluation en formation ? In A. Jorro & Y. Mercier-Brunel (Eds.), *Activité évaluative et accompagnement professionnel* (pp. 99-116). Presses universitaires François-Rabelais. https://doi.org/10.4000/books.pufr.18121

Perrenoud, P. (2000). Mobiliser ses acquis : Où et quand cela s'apprend-il en formation initiale ? De qui est-ce l'affaire ? *Recherche & Formation*, *35*(1), 9-23. https://doi.org/10.3406/refor.2000.1667

Perrenoud, P. (2001). *Développer la pratique réflexive dans le métier d'enseignant professionnalisation et raison pédagogique*. ESF éditeur.

Perrenoud, P. (2010). Et si l'évaluation institutionnelle paralysait le développement professionnel ? In *L'évaluation, levier du développement professionnel ?* (pp. 35-47). De Boeck Supérieur; Cairn.info. https://doi.org/10.3917/dbu.paqua.2010.01.0035

Reuter, Y., Cohen-Azria, C., Daunay, B., Delcambre, I., & Lahanier-Reuter, D. (2013). Évaluation. In *Dictionnaire des concepts fondamentaux des didactiques* (pp. 101-105). De Boeck Supérieur. https://doi.org/10.3917/dbu.reute.2013.01.0101

Schmid, E., & Stalder E., B. (Éds.). (2008). Evaluation de l'étude bernoise LEVA. Pourquoi les jeunes changent de métier durant l'apprentissage. *Panorama*, *1*.

Scriven, M. (1996). Types of evaluation and types of evaluator. *Evaluation Practice*, *17*(2), 151-161.

SERI, 412.101 (2004).

SERI. (2022). *La formation professionnelle en Suisse – Faits et chiffres 2022*. State Secretariat for Education, Research and Innovation. https://www.sbfi.admin.ch/dam/sbfi/fr/dokumente/webshop/2020/bb-f-z-2020.pdf.download.pdf/fakten_zahlen_bb_f.pdf

Seufert, S. (2018). *Flexibilisierung der Berufsbildung im Kontext fortschreitender Digitalisierung*. Secrétariat d'État à la formation, à la recherche et à l'innovation. https://www.sbfi.admin.ch/sbfi/de/home/dienstleistungen/publikationen/publikationsdatenbank/berufsbildung-digitalisierung.html

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. McGraw-Hill Book Company.

Stafford, J., Bodson, P., & Stafford, M.-C. (2006). *L'analyse multivariée avec SPSS*. Presses de l'Université du Québec.

Stern, S., Ehrler, J., Marti, C., & von Stokar, T. (Éds.). (2010). *Evaluation de la formation professionnelle initiale de deux ans (AFP) : Version abrégée*. OFFT.

Tellenbach, D. (2022). Plaidoyer pour une meilleure procédure de qualification : Examens finals dans la formation professionnelle initiale: Une aberration ? *Transfer. Formation professionnelle dans la recherche et la pratique*, *7*(3).

Vial, M. (2012). *L'évaluation située*. De Boeck Supérieur. https://shs.cairn.info/se-reperer-dans-les-modeles-de-l-evaluation--9782804168940-page-347?lang=fr&contenu=resume