Sheri Dion
University of New Hampshire

L2 French Writing Assessment: A Methodological Critique

Abstract

This paper presents a methodological critique of three empirical studies in second language (L2) French writing assessment. To distinguish key themes in French L2 writing assessment, a literature review was conducted resulting in the identification of 27 studies that were categorized into three major themes. The three studies examined in this article each represent one theme respectively. Within this analysis, the underlying constructs being measured are identified, and the strengths and limitations are deliberated. Findings from this detailed examination suggest that three examined studies in L2 French writing assessment have significant methodological flaws that raise questions about the claims being made. From this investigation, several study-specific recommendations are made, and four general recommendations for improving French L2 writing assessment are offered: (1) the social setting in which L2 assessments take place ought to be a consideration (2) the difficulty of tasks and time on task should be taken into account (3) greater consistency should be used when measuring and denoting a specific level of instruction (i.e. "advanced") and (4) universal allusions to "fluency" should be avoided when generalizing one component of L2 competency (such as writing achievement) to other aspects of L2 development.

Key words: French writing, methodological critique, written assessment, language assessment, second language writing assessment

***Journal for Language Teaching*, Volume 53, Number 2, pp. 11 – 25. 2019.**
ISSN 0259-9570. http://dx.doi.org/10.4314/jlt.v53i2.1

Introduction

This paper presents a critique of three empirical studies in second language (L2) French writing assessment. After a brief presentation of the types of articles and relevant themes in L2 writing assessment, the underlying constructs being measured in each study are identified, and strengths and limitations of various assessments are discussed. To demonstrate variation in L2 writing assessments, this analysis draws on a detailed examination of three examples of L2 French writing assessment; however, other studies consistent with the thematic categorization of the literature review are referenced when applicable. In critique of the literature, this discussion demonstrates that the considered studies in L2 French writing assessment have significant methodological flaws that raise questions about the claims being made. Within this examination, several study-specific recommendations for improving French L2 writing assessment are proposed. This analysis concludes with four major recommendations for improving L2 French written assessment: (1) that the social setting in which L2 assessments take place be a variable; (2) the difficulty of tasks and time on task should be taken into consideration; (3) greater consistency should be used when measuring and denoting a specific level of instruction (i.e. "advanced"); and (4) universal allusions to "fluency" should be avoided when generalizing one component of L2 competency (such as writing achievement) to other aspects of L2 development.

Methods

In order to determine which studies were relevant, a keyword search performed using the database Education Resources Information Center (ERIC) for "L2," "French," "assessment," and "written" and again with "L2," "French," "assessment," and writing," identified 31 and 33 peer-reviewed journal articles respectively. The examination of the 64 retrieved articles included article titles, abstracts, keywords, and content to determine their appropriateness for inclusion. From this process, 25 articles met the inclusion criteria of incorporating (1) French language content and (2) written assessment(s) in both the written and writing strands. Within each search, 15 articles met the inclusion criteria from the first search and 10 met the criteria the latter. Two additional articles (Burston & Arispe, 2018; Manchón, 2018) were added from ongoing research, adjusting the total articles examined to 27. Articles were categorized into three themes based on the L2 writing assessment context: (1) technology (T), (2) dimensions (D),[1] and (3) collaborative and self-assessment (C/SA). Empirical studies with a technological component comprise the largest group; 12 of 27 total studies integrate L2 writing assessment with technology. The vast majority of studies (17) are at the university level. A graph depicting these themes in L2 writing assessment is included, see figure 1 (below).

1 The category dimensions (D) was designated for studies in study abroad contexts (e.g. Godfrey, Treacy, & Tarone, 2014), writing portfolios (e.g. Bissoonauth-Bedford & Stace, 2015; Paesani, 2006), and translation (e.g. Cohen & Brooks-Carson, 2001).

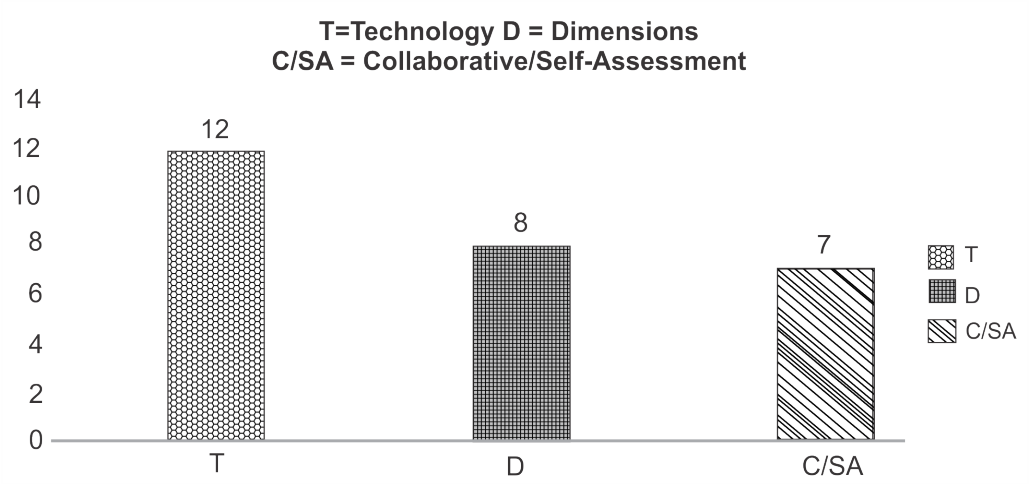


Figure 1. L2 French writing assessment studies by theme

Constructs measuring L2 writing assessment

In terms of how L2 writing is defined, conceptualizations of L2 writing assessment are linked to the linguistic, cognitive, and social dimensions of L2 writing demands (Manchón, 2018). A wide range of underlying theories have guided empirical studies in this area, including but not limited to "linguistic theories, genre theories, theories of literacy development, cognitive models of L1 writing, the problem-solving paradigm in cognitive psychology, research on L2 language learning and language use strategies, and theories of self-regulation" (Manchón, 2018: 249). In addition to theoretical underpinnings of linguistic, cognitive, and social (socio-cultural/socio-cognitive) aspects of L2 writing assessment, assessment in L2 writing involves:

> (a) a whole range of purposes, conditions, and outcomes of writing (i.e. writing in individual/collaborative conditions, with/without the availability of feedback, in diverse time-on-task conditions) (and) b). The various dimensions of writing (i.e. learning to write/writing to learn content/writing to learn language. (Manchón, 2018: 259-260)

Recognizing the multidimensional nature of L2 writing assessment, an examination of empirical research indicated, unsurprisingly, constructs in which no singular definition of L2 writing assessment holds. Constructs measuring L2 writing proficiency ought not to be measured subjectively, yet in some cases (Arens & Jansen, 2016; Dagenais, Toohey, Fox, & Singh, 2016; Gabaudan, 2016), definitions of constructs are neither accompanied with clear descriptions nor are they complemented by objective results. This presents several challenges to measurement and has implications for validity and reliability, attempts to draw inferences from data, and possible generalizability.

More precisely, the term validity refers to "the approximate truth of an inference"[2] (Shadish, Cook, & Campbell, 2002: 34). Typically, validity is employed as a judgment about whether or not evidence from empirical findings supports an inference as being true or correct (Shadish et al, 2002). There are several types of errors and threats to validity that can affect inferences that are made, and when explicitly referenced in this article, an explanation of the various validity typologies will be offered. In addition, reliability refers to consistency and should be assessed and reported when measuring constructs. Among the many remedies for unreliability in measurement, some examples include increasing the number of measurements and improving the quality of measures (Shadish et al, 2002). In addition, a treatment that is inconsistently applied can affect conclusions about covariation, yet there are several ways to improve, measure, and analyze treatment implementation in order to reduce this threat (Shadish et al, 2002). A more detailed analysis of variance in construct definition, strengths and limitations of each study, and study-specific recommendations are presented below.

Thematic empirical studies

A single study from each of the three identified categories was chosen to present an in-depth analysis of methodology, procedures, and implications for L2 French written assessment. Although each of the studies included this investigation would have been equally likely to contribute to this discussion, the three examples that were chosen were selected at random. Due to the limitations of space and time, a discussion of each of the 27 articles is outside the scope of the present article. However, when possible, additional articles identified in the literature review are noted in support of the recommendations made for improving L2 written assessment.

L2 French written assessment and technology

Studies in L2 written assessment and technology include examples such as the pros and cons of virtual worlds (Garrido-Iñigo & Rodríguez-Moreno, 2013) and an examination of how iPads might be assets in L2 learning (Dagenais et al, 2016; Pellerin, 2014). Granfeldt and Agren's (2014) research was chosen to illustrate the theme of L2 writing assessment with technology. This study presents a mixed-methods approach to online writing evaluation that implemented a diagnostic assessment tool called Direkt Profil[3] and compared its automated analysis with L2 teacher's two-part assessment (1) to make separate assessments of comments and form on a Likert scale (0-6) and (2) to comment on distinctive features of the written L2 texts being assessed.

2 An inference can also be referred to as a knowledge claim or proposition (Shadish et al, 2002).

3 Direkt Profil is available online and free. Retrieved from http://profil.sol.lu.se/profil/logon.jsp

Direk Profil is an interlanguage diagnostic tool (technically a parser)[4] for L2 French that "provides immediate and detailed feedback indicating how certain types of linguistic structures, correct or incorrect, are related to different stages of development" (Granfeldt & Agren, 2014: 285).

The goal of this research was not to measure student achievement; this study involved a profile analysis of students' linguistic abilities, grouping students into levels of L2 developmental stages (0-6, with 0 being beginner and 6 being the most advanced, according to Bartning and Schlyter's (2004) criteria[5] of stages of development). The sample population in this study involved 100 L2 secondary Swedish students who wrote 400 L2 French narratives about picture stories included in the student's curriculum: "Le voyage en Italie"[6] ("The trip to Italy"). Upon closer inspection of an example of this picture story and Bartning and Schlyter's (2004) criteria, an identifiable threat to content validity[7] seems apparent: secondary L2 students' descriptions of this relatively simplistic, beginner-level story implies a need for explicit instructions about which content to include in written narratives to attain the highest levels (5-6, inclusive of the subjunctive and conditional phrases). Further, students were grouped into linguistic ability (i.e. generalizes their "fluency") based on one instance of writing achievement which, with the exception of teacher comments, was based solely on the accuracy of lexical and morphological French. In other words, neither the abilities to synthesize and analyze a story nor the ability to perform additional tasks incorporating other L2 French skills were involved in the creation of student groups.

The constructs measuring L2 writing assessment in this context were approximately 25 different phenomena, including but not limited to: "finiteness, tense, aspect, verbal agreement, negation, subordination, number and gender agreement, subject and object pronouns," (Granfeldt & Agren, 2014: 287). A critique of the constructs being measured in this study is that there was not only a lack of uniformity as to what exactly is being measured, there is also the question of how it was being measured, i.e. "no special instructions about any particular criteria preceded the teachers' assessments" (Granfeldt & Agren, 2014: 288). Inter-rater reliability was also not addressed. In addition, 25 different aspects of lexical and morphological French (without regard to the level of difficulty) seems beyond the scope of what secondary school students, might include in short, written narratives. This study also did not take into account the social context and time on task during which the work was completed.

4 Which can be defined as an interpreter that breaks data into smaller elements. It takes input in the form of a sequence or program instructions to build a data structure. Direct Profil is a "partial" parser, meaning it is designed to analysis phrases/groups rather than whole sentences. Retrieved from https://www.techopedia.com/definition/3854/parser

5 Retrieved from https://www.cambridge.org/core/services/aop-cambridge-core/content/view/E51B5BDB4ABEE63E6CFF9566870472BB/S0959269504001802a.pdf/itineraires_acquisitionnels_et_stades_de_developpement_en_francais_l2.pdf

6 An example of one of these stories is included, see Appendix (p. 18).

7 Content validity refers to the extent to which a tool measures the intended area of content (McGoey, Cowan, Rumrill, & LaVogue, 2010).

An advantage of using the Direkt Profil database was the immediate and detailed feedback produced in addition to the feedback provided by the seven teachers in this study. To the point of the nature of Direkt Profil's feedback, the results showed language form to be a better predictor than content in overall assessment. This finding offers one potentially useful way of integrating "the automated diagnosis of developmental stage as part of the assessment of learners' language abilities" (Granfeldt & Agren, 2014: 298). However, to the point of using a partial parser,[8] Direkt Profil is not capable of assessing content, which calls into question how this conclusion can be made based primarily on seven subjective, independent teacher comments. Further, the jump to endorse this technological tool without weighing the drawbacks more holistically seems to negate other components in this study.

For example, qualitative analyses revealed more agreement among the teachers assessing student work. This means that across levels, the teachers' assessments of students' work were more consistent, and when comparing teachers' results with those of Direkt Profil, Direkt Profil's assessment showed some discrepancies. The majority of variation in teachers' assessments was observed primarily in the intermediate levels, yet discrepancies in Direkt Profil's assessment was observed at almost all levels of student written assessment. In terms of using an automated profile analysis to interpret results, it seems that some insight is valuable yet this diagnostic "cannot replace teacher's evaluation of strengths and weaknesses in learner production and their constructive feedback to individual learners" (Granfeldt & Agren, 2014: 303). In this regard, if the assessment of L2 writing process necessitates teacher feedback as opposed to reliance on a diagnostic tool, this tool would need to be more explicitly and clearly validated in order to justify its use and implementation.

Among additional limitations, while it is certainly an interesting exploration of technology, Direkt Profil had several challenges in practical use. The reliability of Direkt Profil was challenged by incorrect spelling and unrecognizable or non-programmatic grammatical relationships. These two specific challenges resulted in "lower precision and recall scores" (Granfeldt & Agren, 2014: 292-293), and resulted in the need to add specific structures and additional analyses of spelling. However, the additional structures and spelling analyses were added after the investigation had already begun. This particular challenge relates to unreliability of treatment implementation, a threat to statistical conclusion validity, since the treatment "intended to be implemented in a standardized manner is implemented only partially for some respondents, effects may be underestimated compared with full implementation" (Shadish et al, 2002: 45). For the qualitative teacher assessments, bias on the part of the teachers' subjectivity in grading the narratives was not addressed, nor was reactivity to the setting or individuals influencing this study. A goal in qualitative work is "not to eliminate this influence, but to understand it and use it productively" (Maxwell, 2013: 125), yet recognition of these threats is missing from this study.

8 Again, a partial parser is an interpreter that breaks data into smaller elements. It takes input in the form of a sequence or program instructions to build a data structure. In this sense, a partial parser analyzes specific phrases rather than entire sentences.

Recommendations specific to this study include (1) the need to pilot Direkt Profil before implementing its use, (2) provide a rationale for performing this work, (3) training for teachers assessing this work, (4) inter-rater reliability checks, and (5) the need to use a well-established, objective measuring tool to categorize student L2 written level of "fluency." Between seven subjective teacher reports in comparison with a flawed diagnostic tool with no capacity to comment on content, the categorization of student written work may also have suffered from additional unidentified discrepancies.

Dimensions: Rasch Measurement Theory in Examination of French Grammar

In Vogel and Englehard's (2011) study, the authors examine whether guided inductive vs. deductive instruction of French grammar has a relationship with student achievement through the lens of Rasch Measurement Theory (RMT).[9] This study examines the effects of two instructional approaches to teaching French grammar, guided inductive and deductive approaches (predictors) on student achievement (outcome) measured on identical pre- and post-test evaluations. Two many-faceted Rasch (MFR) models[10] were used to evaluate student learning of French grammar: the first to assess student learning from pre- to post-test and the second to examine the effects of the inductive and deductive instructional approaches. In this study, L2 written achievement was defined as knowledge of 10 basic grammatical structures at the university-level intermediate French. These structures were measured using 1.) identical pre and post-tests and 2.) four-question open-ended written quizzes administered at the end of each of the 10 lessons with a score of 0, 1, or 2. Reported findings suggest that student achievement on French grammatical structures, on average, is higher when taught with an inductive approach.

Strengths of this study include the integration of RMT to address student performance over time, item difficulty, and instructional approach. Among limitations, students took identical pre- and post-tests. This represents an internal validity testing threat, meaning that "exposure to a test can affect test scores on subsequent exposures to that test, an occurrence that can be confused with a treatment effect" (Shadish et al, 2002: 55). A second critique involves the primary investigator's role in teaching all 10 lessons to both sections which presents an experimenter expectancy threat to construct validity.

9 Rasch Measurement Theory is a psychometric theory that was developed to improve the "precision with which researchers construct instruments, monitor instrument quality, and compute respondents' performances" (Boone, 2016: 1). The Rasch model is considered an advanced measurement approach that addresses limitations such as lack of control for difficulty level of scale items. In this sense, the Rasch model "provides better measurement of items and more precise measurement of scale" (Brinthaupt & Kang, 2014: 242).

10 More specifically, a many-faceted Rasch model allows for an estimate of subscale difficulties through Rasch calibration, accounting for both item difficulty and person ability (Brinthaupt & Kang, 2014).

This particular threat has been identified since the "experimenter can influence participant responses by conveying expectations about desirable responses, (and) those expectations are part of the treatment construct actually tested" (Shadish et al, 2002: 73). To address this threat, a few of the lessons were videotaped to ensure the primary investigator did not favor one approach over the other. However, an exact definition of "a few" is not offered, and 20 total lessons were administered, which is substantial.

An additional limitation of this study involves a significant history threat[11] to internal validity, indicating that instruction concurrent with the treatment may have contributed to the observed effect. The three regular classroom instructors who otherwise taught the classes in this study were "specifically asked not to cover the 10 structures investigated during class time" (Vogel & Englehard, 2011: 272). However, this study occurred over the course of an entire fall term, featuring the pre-test at the start and post-test at the end of the term. Given that the investigators were not present other than for their weekly lessons and each section met and received French instruction during an additional three to four classes per week throughout the term, this unobserved instructional time is considerable. Although teachers were asked not to teach certain content, there certainly would be overflow of the basic structures in question (e.g. examples in the course text, assignments, evaluations, etc.) even if those structures were not explicitly taught.

A treatment diffusion threat to construct validity represents a final identified limitation in this study. The participants may have received "services from a condition to which they were not assigned, making construct descriptions of both conditions more difficult" (Shadish et al, 2002: 73). The authors chose a within-participants design instead of random assignment. Doing so allows each student to serve as his or her control and to reduce rival hypotheses, although explanation of why this was chosen vs. random assignment was not clearly stated. The first 10 target structures were taught inductively to the first group and deductively to the second, and then each subsequent grammatical structure (10 additional) taught alternated between a deductive and inductive presentation. In this design, the inductive group would therefore receive 15 inductive and five deductive lessons, and the deductive group would receive the reverse. Although this design has benefits, it seems that carryover effects may cause problems. The authors minimally refer to this possibility, yet do not provide further rationale other than, "there are advantages of this design, (and) there are limitations such as carryover effects" (Vogel & Englehard, 2011: 272).

Finally, content validity, referring "to the extent to which a measurement instrument reflects the intended area or domain of content," (McGoey, Cowan, Rumrill, & LaVogue, 2010: 109) can be addressed. This study incorporates many relevant and appropriate grammatical concepts, yet it seems to almost intentionally avoid others that might have been more reflective of an intermediate level.

11 A history threat to internal validity represents one reason why inferences about a relationship between two variables may be incorrect and means that "events occurring concurrently with treatment could cause the observed effect" (Shadish et al, 2002: 55).

Otherwise put, this study examines relatively elementary content with intermediate students and seems to omit concepts which necessitate some reflection or evidence of acquiring intermediate L2 skills. For example, the authors include a few (4 of 16) verbs of the passé composé with être (that are all regular conjugations), but not avoir, and they do not ask students to make a choice between avoir/être. In another instance, they ask students to choose between the subjunctive or the infinitive but not the indicative (present) tense, which would have been a far more difficult task.

Specific recommendations for this study include the need for a peer reviewer be present in the room during all lessons to verify that one instructional approach was not favored over the other. An alternative way to design the groups would be to have three groups, one control, one deductive, and one inductive approach, although this design may introduce additional rival hypotheses and may require more than one peer reviewer. Also, both the inductive/deductive groups were given the exact same review PowerPoint slideshow at the end of the unit, representing an example of inductive instruction (and highly favored visual learners). As one additional problem, students would "chorally answer" (Vogel & Englehard, 2011: 272) the PowerPoint review questions which can allow students who do not understand to proceed oftentimes unnoticed and does not offer ample time for reflection. The authors should consider revising this review strategy, which also points to an overall weighted benefit to inductive instruction and is consistent with their findings.

Collaborative and Self-Assessment: 9th Grade Self-Assessment

In a self-assessment context, Van Reybroeck, Penneman, Vidick, and Galand's (2017) quasi-experimental study involved a pre/post-test to observe eight lessons with four treatment groups of ninth graders and a control group (126 total students) at two schools to determine which and if various interventions had an effect on L2 grammatical spelling of past participles. Each student group received, respectively: "progressive treatment alone, coupled with self-assessment, coupled with feedback, or coupled with self-assessment and feedback" (Van Reybroeck et al, 2017: 1965). Progressive treatment was defined as a treatment that enables them to automate grammatical rules that "includes exercises to increase the cognitive load throughout the intervention" (Van Reybroeck et al, 2017: 1968). Otherwise put, the exercises become progressively more sophisticated as they add grammatical concepts. The construct defining L2 writing assessment in this study was the accurate spelling of French past participles, measured in a pre- and post-exam (50 minutes each; 150 minutes total). Reliability was addressed with twice-weekly meetings between the two French instructors and the researchers to ensure similar implementation between schools. In these observations, students receiving self-assessment coupled with feedback were juxtaposed against a control group of thirty-six students with standard instruction alone (Van Reybroeck et al, 2017).

Strengths of this study include the consideration of context in which the study took place: of the two schools investigated, "one school was of low socio-economic status whereas the other school was in an area with middle socio-economic status" (Van Reybroeck et al, 2017: 1971). Other promising aspects of this study are its aim to investigate which

teaching practices have a relationship on student achievement, here defined as the correct reproduction of French past participles, and its singular focus on one component of written assessment. Results were in favor of adding a self-assessment component and suggested that this group "improved even more on spelling tests, including free text production" (Van Reybroeck et al, 2017: 1966). The self-assessment component of this study was also hypothesized as having a potential impact on increasing students' motivation.

In terms of limitations in this study, the assumption that 9th graders can accurately measure their own language proficiency is questionable. However, their self-assessments in this context were in juxtaposition with purported objectively graded results. Further, in quasi-experiments, "control groups may differ from the treatment group in many systematic (non-random) ways other than the presence of the treatment" (Shadish et al, 2002: 14). For example, the control group, in particular in the low-SES school may have included students who were disadvantaged and had less access to resources. Additionally, the regular classroom teachers in this study were also the administrators of eight weekly lessons in two months. This could pose additional threats: an experimenter expectancies threat to construct validity and internal threats to history and testing, as regular classroom instruction and other forms of homework, review, and assessment could have an influence on the observed effect (Shadish et al, 2002: 55, 73). Finally, the pre-test was administered three weeks before winter break, indicating that three weeks' instruction continued before the treatment began, representing ambiguous temporal precedence or a "lack of clarity about which variable occurred first" (Shadish et al, 2002: 55). Suggestions for this study include (1) classroom observations of the administered lessons in addition to twice-weekly teacher meetings (2) administering the pre-test after the winter break, before instruction starts, and not three weeks prior, and (3) gathering student feedback on the self-assessment component in the form of a questionnaire or meeting with students in those groups.

Conclusion

Based on this critique, four main recommendations are offered for future research in L2 writing assessment. (1) Given the relevance of context in which the L2 writing assessments are studied, it is recommended that the social setting (i.e. class, school, community) in which the L2 writing assessments are examined be included as a variable. This recommendation is supported with evidence from additional studies (Gabaudan, 2016; Van Reybroeck et al, 2017). (2) The difficulty of tasks and time on task should be taken into consideration. In addition to Vogel and Englehard's (2011) study, these concerns are also evident in additional studies examined in the literature review (Caws, 2006; Godfrey, Treacy & Tarone, 2014). (3) More consistent measures should be used when comparing levels of instruction (Arens & Jansen, 2016; Dagenais et al, 2016). (4) Generic allusions to "fluency" or "proficiency" should be avoided when generalizing one component of L2 competency (such as writing achievement) or other allusions to it being congruent with other aspects of L2 development (Burston & Arispe, 2018; Manchón,

2018). Further, the uncritical designation of technological components as a potential benefit to students in L2 writing assessment is also problematic and was evident in the literature (Meara, Rodgers, & Jacobs, 2000; Nicol, 2009; Pellerin, 2014). While L2 written proficiency can be measured both quantitatively and qualitatively, it is also suggested that an objective guideline such as those used by the American Council on Teaching of Foreign Languages (ACTFL, 2012) also be taken into consideration. As a final suggestion, writing development in L2 relates to writing ability in L1, the consideration of cross-curricular examination of student development in writing is not included in the scope of this paper, yet such considerations ought to be taken into account as they relate to the social and cognitive development of students.

[The author would like to acknowledge Dr. Suzanne E. Graham for her excellent instruction, mentorship, and guidance in quantitative methods.]

References

American Council of the Teaching of Foreign Languages. 2012. Retrieved from https://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf [Accessed: 12 October 2019]

Arens, A. K., & Jansen, M. 2016. Self-concepts in reading, writing, listening, and speaking: A multidimensional and hierarchical structure and its generalizability across native and foreign languages. *Journal of Educational Psychology.* 108 (5): 646-664. DOI: 10.1037/edu0000081

Bartning, I., & Schlyter, S. 2004. Itinéraires acquisitionnels et stades de développement en français L2. *Journal of French Language Studies*. 14: 281-300. DOI: 10.1017/S0959269504001802

Bissoonauth-Bedford, A., & Stace, R. 2015. Building a Writing Community Through Learning of French. *Journal of University Teaching and Learning Practice*. 12 (2): 1-22. Available at: https://ro.uow.edu.au/jutlp/vol12/iss2/7 [Accessed: 13 October 2019]

Boone, W. J. 2016. Rasch Analysis for Instrument Development: Why, When, and How? *Cbe - Life Sciences Education.* 15 (4): 1-7. DOI: 10.1187/cbe.16-04-0148

Brinthaupt, T. M., & Kang, M. 2014. Many-Faceted Rasch Calibration: An Example Using the Self-Talk Scale. *Assessment.* 21 (2): 241-249. DOI: 10.1177/1073191112446653

Burston, J., & Arispe, K. 2018. Looking for a needle in a haystack: CALL and advanced language proficiency. *Calico Journal.* 35 (1): 77-102. DOI: 10.1558/cj.31594

Dagenais, D., Toohey, K., Fox, A. B., & Singh, A. 2016. Multilingual and multimodal composition at school: *ScribJab* in action. *Language and Education.* 31 (3): 263-282. DOI: 10.1080/09500782.2016.1261893

Caws, C. 2006.Assessing Group Interactions Online: Students' Perspectives. *Journal of Learning Design.* 1 (3): 19-28. DOI: 10.5204/jld.v1i3.23

Cohen, A. D., & Brooks-Carson, A. 2001. Research on direct versus translated writing: Students' strategies and their results. *Modern Language Journal.* 85 (2): 169-188. DOI: 10.1111/0026-7902.00103

Gabaudan, O. 2016. Too soon to fly the coop? Online journaling to support students' learning during their Erasmus study visit. *Recall.* 28 (2): 123-146. DOI: 10.1017/s0958344015000270

Garrido-Iñigo, P., & Rodríguez-Moreno, F. 2013. The Reality of Virtual Worlds: Pros and Cons of Their Application to Foreign Language Teaching. *Interactive Learning Environments.* 23 (4): 453-470. DOI: 10.1080/10494820.2013.788034

Godfrey, L., Treacy, C., & Tarone, E. 2014. Change in French second language writing in study Abroad and domestic contexts. *Foreign Language Annals*. 47 (1): 48-65. DOI: 10.1111/flan.12072

Granfeldt, J., & Agren, M. 2014. SLA developmental stages and teachers' assessment of written French: Exploring Direkt Profil as a diagnostic assessment tool. *Language Testing.* 31 (3): 285-305. DOI: 10.1177/0265532214526178

Manchón, R. M. 2018. Past and future research agendas on writing strategies: Conceptualizations, inquiry methods, and research findings. *Studies in Second Language Learning and Teaching: Ssllt.* 8 (2): 247-267. DOI: 10.1515/9781614511335-003

Maxwell, J. A. 2013. *Qualitative research design: An interactive approach*. Thousand Oaks, Calif: SAGE Publications. DOI: 10.1093/obo/9780199756810-0126

McGoey, K. E., Cowan, R. J., Rumrill, P. P., & LaVogue, C. 2010. Understanding the psychometric properties of reliability and validity in assessment. *Work*. 36: 105–111. DOI: 10.3233/WOR-2010-1012

Meara, P., Rodgers, C., & Jacobs, G. 2000. Vocabulary and neural networks in the computational assessment of texts written by second-language learners. *System.* 28 (3): 345-354. DOI: 10.1016/s0346-251x(00)00016-6

Nicol, D. 2009. Assessment for learner self-regulation: enhancing achievement in the first year using learning technologies. *Assessment and Evaluation in Higher Education.* 34 (3): 335-352. DOI: 10.1080/02602930802255139

Paesani, K. 2006. Developing Literacies - Exercices de style: Developing Multiple Competencies Through a Writing Portfolio. *Foreign Language Annals.* 39 (4): 618-639. DOI: 10.1111/j.1944-9720.2006.tb02280.x

Pellerin, M. 2014. Using Mobile Technologies with Young Language Learners to Support and Promote Oral Language Production. *International Journal of Computer-Assisted Language Learning and Teaching.* 4 (4): 14-28. DOI: 10.4018/978-1-5225-7663-1.ch034

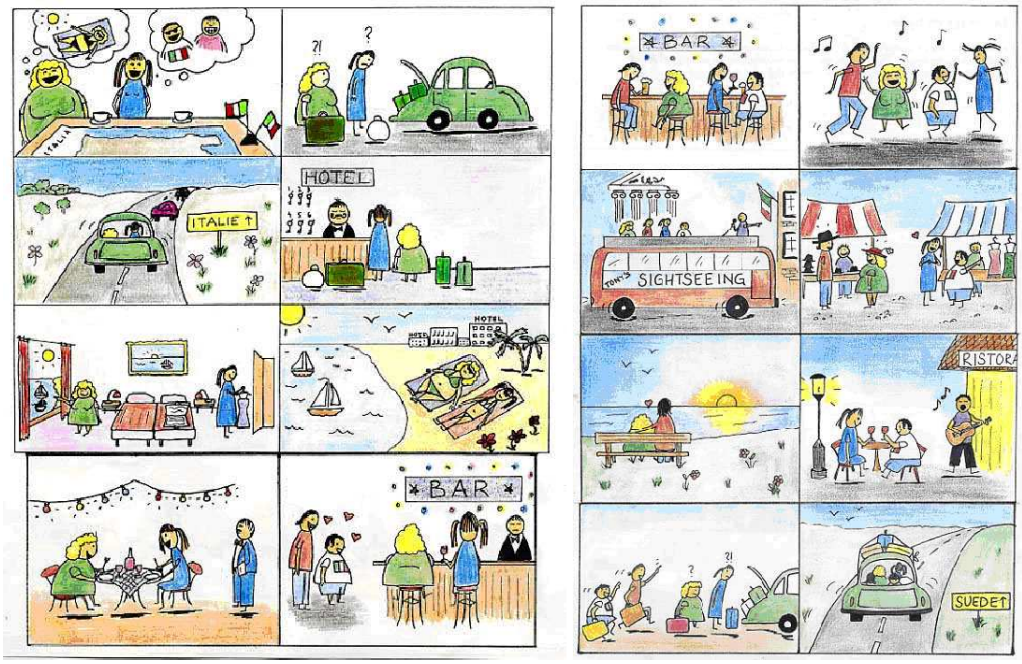
Shadish, W. R., Cook, T. D., & Campbell, D. T. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth Cengage Learning.

Van Reybroeck, M., Penneman, J., Vidick, C., & Galand, B. 2017. Progressive treatment and self-assessment: effects on students' automatisation of grammatical spelling and self-efficacy beliefs. *Reading and Writing: an Interdisciplinary Journal.* 30 (9): 1965-1985. DOI: 10.1007/s11145-017-9761-1

Vogel, S. P., & Engelhard, G. J. 2011. Using Rasch Measurement Theory to Examine Two Instructional Approaches for Teaching and Learning of French Grammar. *Journal of Educational Research.* 104 (4): 267-282. DOI: 10.1080/00220671003733815

Appendix

“Le voyage en Italie” (Granfeldt & Agren, 2014)



ABOUT THE AUTHOR(S)

Sheri K. Dion

Affiliation: University of New Hampshire, USA
Email: sheridion@gmail.com

Sheri K. Dion is a Ph.D. Candidate at the University of New Hampshire. Her research centers on intersectional dynamics in language pedagogy, quantitative and qualitative methodologies, second language writing assessment, and curriculum theory. She has recently published on culture and diversity in language teaching and the promises and constraints of cosmopolitanism as an educational project. Prior to and throughout her doctoral studies, Ms. Dion has been a French instructor at the secondary level for many years.