The development and validation of an intercultural competence evaluation instrument for upper secondary foreign language teaching

ESTHER SCHAT [a]

EWOUT VAN DER KNAAP [b]

RICK DE GRAAFF [c]

[a] *Utrecht University, The Netherlands*
e.schat1@uu.nl

[b] *Utrecht University, The Netherlands*
e.w.vanderknaap@uu.nl

[c] *Utrecht University, The Netherlands*
r.degraaff@uu.nl

Abstract

Intercultural competence is a crucial element of foreign language education, yet the multifaceted nature of this construct makes it inherently difficult to assess. Although several tools for evaluating intercultural competence currently exist, research on their use in secondary school settings is scarce. This study reports on the development and validation of an instrument intended specifically for use in foreign language literature education. To this end, tangible learning objectives for intercultural literary competence were developed based on five dimensions of intercultural communicative competence (Byram, 1997) and six categories of literary competence (Witte, 2008). These objectives were re-formulated for implementation in a student questionnaire. Using exploratory and confirmatory factor analysis, the construct validity of the instrument was investigated among a sample of 164 secondary school students in the upper forms (aged 16-19) of pre-university education in the Netherlands. Although the results supported the hypothesized two second-order factor structure, the model fit indices were less favourable compared to the fit indices of an alternative five first-order factor model. Subsequently, correlation and summability analyses were performed to test the reliability of the instrument. Future research and implications are discussed.

Introduction

The ability to understand and communicate with people across all kinds of cultural divisions is indispensable in today's transcultural and globalized societies. There is, thus, a clear imperative to prioritize intercultural competence as a main objective in secondary education (Council of Europe, 2018), preparing adolescents for "interaction with people of other cultural backgrounds, teaching them skills and attitudes as well as knowledge" (Byram & Wagner, 2018, p. 140). Research in the field of foreign language education has highlighted the benefits of literature-based pedagogies for developing intercultural competence, with recent publications describing recommendations for pedagogical practices and underlining the contribution of literature education to intercultural development (e.g., Matos & Melo-Pfeifer, 2020; Porto & Zembylas, 2020). However, validated instruments that monitor such development and evaluate the effectiveness of pedagogical practices are absent at the secondary level. As language teachers must both evaluate their programs and assess whether students are actually improving, tools that map students' intercultural development through foreign language literature education are highly needed. Moreover, given how commonly foreign languages are taught in secondary schools around the world, developing such tools is also urgent as its backwash effect "is crucial in ensuring that learners and teachers pay serious attention to intercultural competence and include it in a systematic way in their planning and implementation of the curriculum" (Barrett *et al.*, 2014, p. 19). The aim of the present study is to develop and validate an instrument to measure the intercultural competence of secondary school students in foreign language literature teaching.

Background and context

In order to design and implement such an instrument, we performed our study in the context of foreign language education in the upper forms of pre-university education in the Netherlands. In the Dutch secondary school setting, literature has always been an obligatory component of the foreign language curriculum, but an explicit intercultural perspective has been absent so far. The national core standards for literature include attainment targets for reading for personal development and for knowledge of literary history, concepts and terminology (Meijer & Fasoglio, 2007) but do not mention intercultural aspects of foreign language literature teaching. Teachers enjoy a great amount of freedom in how to implement the three standards, and recent studies on literature education in the Netherlands have revealed that both foreign language teachers (Lehrner-te Lindert, 2020) and their students (Bloemert *et al.*, 2019) find the cultural element of literature teaching most important. However, intercultural literature teaching is still in its infancy. A survey on the topic has shown that literary texts in educational practice tend to be reduced to an informative representation of the national culture/s associated with the foreign language, or as a source for language proficiency due to the absence of guidelines, materials and tools for a more in-depth and visible intercultural approach (Schat *et al.*, 2018). Literature as a tool for intercultural understanding is gaining territory in the Dutch curricular debate (Meesterschapsteam MVT, 2018) and, recently, major governmental plans have started to revise the national curriculum in which intercultural competence will most likely have a more prominent position (Curriculum.nu, 2019). Nevertheless, at the time of writing Dutch teachers lack objectives and evaluation instruments for intercultural literature teaching.

On an international scale, tangible learning objectives for intercultural literature teaching that offer guidance to secondary school teachers are scarce. Recently, a theoretical model of the intercultural reader (Hoff, 2016) specifically aimed at the secondary level was developed, yet it does not address concrete learning objectives. In an earlier study, Burwitz-Melzer (2001), presented substantial objectives with associated behaviors for secondary school students, but this formulation lacked clarity with respect to how these related to the several dimensions of intercultural competence. The

recent compendium to the CEFR (Council of Europe, 2018) does not provide direction either. Although it does contain descriptors for "analysis and criticism of creative texts" and "a personal response to creative texts," it does not provide cultural descriptors for literature (Alter & Ratheiser, 2019). Moreover, there is a more general call for intervention studies on the topic of literature teaching with an empirical approach (Schrijvers *et al.*, 2019). Increasing the availability of assessment instruments based on sound operationalization of descriptors, and with demonstrated reliability and validity, may help researchers to set up well-designed intervention studies. Thus, in order to implement intercultural education more firmly in secondary foreign language curricula, while simultaneously contributing to the current academic discourse on the added value of literature education, the importance of developing an evaluation instrument for literature teaching cannot be overstated.

A possible explanation for the lack of tools to evaluate intercultural progress through foreign language literature education at the secondary level might be due to the general difficulties associated with assessing intercultural competence. Since intercultural competence encompasses attitudes, knowledge, and skills, the multi-dimensionality of the construct as an affective, cognitive and behavioral ability imposes obstacles on educators in terms of assessment (Fantini, 2020; Van de Vijver & Leung, 2009). As attitudes are not directly observable, it follows that evaluations by others are debatable. From an ethical point of view, it is also questioned whether it is desirable to let others assess personal traits (Hoff, 2020). Self-reports are, therefore, often suggested as more responsible assessment instruments. However, given that intercultural competence is frequently understood in terms of the individual's ability to communicate and behave effectively and appropriately, it is argued that this type of assessment experiences validity problems because appropriateness "can only be measured through others' perspectives" (Deardorff, 2016, p. 122). Regarding the reliability of self-reports, it has also been highlighted that respondents who give themselves a high score may not have carried out in-depth reflection, and that maximum scores on questionnaires are meaningless as intercultural competence is a lifelong developmental process (Hoff, 2020; Sercu, 2004). Consequently, to ensure valid and reliable evaluations of intercultural development, it is strongly recommended that multiple measures in which indirect assessment (e.g. self-reports, portfolio) is combined with direct assessment (e.g. tests, observations) are used rather than singular measures.

Another possible explanation for the lack of tools available to map intercultural development through literature education is the poor adaptability of intercultural competence instruments across different educational contexts. While a self-reporting scale has been developed for the context of foreign language literature teaching at higher education (Rodríguez & Puyal, 2012), this scale is of no purpose in other contexts as it alludes to how certain elements of specific literary texts had contributed to intercultural development. Research on the topic of intercultural competence assessment reveals that there are few instruments that researchers or teachers can use without further validation or revision because theoretical and operational definitions of the construct vary greatly across contexts (Deardorff, 2016; Perry & Southwell, 2011). Moreover, considering the abundance of conceptualizations of intercultural competence (Spitzberg & Changnon, 2009) and the diverse contexts in which its evaluation is essential, evaluation instruments must always represent the conceptual and contextual specifics of the construct. Concerning this matter, Deardorff (2016, p. 121) argued that "the starting point should not be to select a measurement tool. Rather, it should be to clarify what specifically is to be assessed by defining terminology based on research and existing literature, and then developing specific goals and measurable objectives based on those definitions." In other words, in order to examine or develop appropriate measurement tools, it is essential to, first, define intercultural competence in the specific educational context, and second, to formulate tangible learning objectives.

Theoretical framework: Intercultural literary competence

Following the reasoning above, it is necessary, firstly, to clarify how intercultural competence is defined within the context of foreign language literature education at the secondary level and, secondly, to describe its specific learning objectives. We approach literary pedagogy from a humanistic paradigm where education serves to encourage students to reflect on the world, on themselves and on others (Biesta, 2006; Nussbaum, 1998). Much research on the topic of literary texts in foreign language education points to dialogic teaching to foster this kind of reflection (Bredella, 2008; Delanoy, 2005; Kramsch, 1993). In dialogic literature education two types of dialogue are promoted: a dialogue with the text and a dialogue with others. Students reflect on the Self through an internal dialogue with the text, becoming aware of their own thoughts, feelings and ideas, and it is through external dialogues with peers about their personal responses to the text, that students explore otherness (Schrijvers *et al.*, 2019). In other words, a dialogic approach encourages students to communicate with and about literature, focusing on personal responses instead of correct answers. With the humanistic aim of gaining relevant insights into the Self and Other, the promotion of both types of dialogue is essential as it fosters an open attitude to different perspectives and opinions. As such, dialogic literature education and intercultural competence emerge as two strongly intertwined entities, as the latter entails looking at the relationships between one's own cultures and other cultures to cultivate an open attitude and a deeper understanding of others. With reference to dialogic teaching for intercultural competence, Bredella (2008, p. 12) emphasizes how a dialogue with the text presents students with possibilities "to evaluate the world of the text and relate it to their own." Kramsch (1993), on the other hand, stresses the dialogue with other readers, arguing that through communication about possible contrasting interpretations of a literary text with peers, students gain insight into their own cultural assumptions and that of others. Delanoy (2005) proposes a more socio-cultural approach to literature teaching, in which students are encouraged to critically question the ideological meaning making of literary texts.

While the above mentioned research on the topic of interculturality and dialogic literary pedagogy is mainly theoretical and focuses exclusively on higher education, operationalizations of what dialogic intercultural literature teaching actually implies for secondary school students in a foreign language class are scarce. We have, therefore, in an earlier study (Schat *et al.*, 2020), described the construct of intercultural literary competence (ILC) in the context of foreign language teaching at secondary schools. As shown in Figure 1, we started from the commonly used model of intercultural communicative competence (Byram, 1997) and reformulated the five *savoirs* (attitudes, knowledge, skills of interpreting and relating, skills of discovery and interaction and a critical cultural awareness). Although Byram's model has been critiqued due to its perceived emphasis on national cultures (Risager, 2007; Hoff, 2020), we selected the original model for our purposes considering its strong focus on objectives. In order to make the five *savoirs* specific for literature teaching at the secondary level, Witte's (2008) literary competence model, a widely used framework in secondary education in the Netherlands, was used. We selected this particular literary competence model as it distinguishes, in line with dialogic literature teaching, "communication with literature" and "communication about literature." It outlines four categories for communication with literature: 1) willingness to make an effort, 2) understand content and meaning, 3) recognize the text's internal characteristics and 4) make connections between texts and the external world. For communication about literature two categories are described: 1) individual and 2) with others (Witte, 2008, p. 176).

To describe the construct of ILC, the six categories of literary competence were integrated into the five *savoirs*. As can be seen in Figure 2, the four categories of communication with literature were integrated into Byram's dimensions of attitude, knowledge, and skills of interpreting and relating, and the two categories of communication about literature were merged with the dimensions of skills of discovery and interaction and a critical cultural awareness. As such, ILC implies that students,

through dialogue with a foreign language text, can become curious and open to other cultures and expand their cognitive knowledge about cultures in general, and use that information to explain and relate a text. Dialogue with others about a foreign language literary text should incite the students' ability to use the foreign language in order to gain insight into the cultural frameworks of peers and to evaluate a text critically. Along the five dimensions learning objectives were formulated, adhering as close as possible to the objectives of Byram (1997).

Dialogue with literature	**Attitudes** (*savoir-être*): the extent to which the student is willing to approach representations of other cultures and one's own in the foreign language literary text with an open and curious attitude, and to suspend disbelief about other cultures and one's own.
	Knowledge (*savoirs*): the extent to which the student can use the foreign language literary text to expand one's knowledge of social groups and their products and practices in one's own and in other cultures, and the general processes of societal and individual interaction.
	Skills of interpreting and relating (*savoir comprendre*): the extent to which the student can recognise textual elements in the foreign language literary text, and the extent to which the student can explain the text and relate it to documents or events from one's own.
Dialogue about literature	**Skills of discovery and interaction** (*savoir apprendre / faire*): the extent to which the student can use the foreign language to acquire new knowledge (of a culture and cultural practices) and operate knowledge, attitudes and skills under the constraint of real-time communication and interaction about the foreign language literary text.
	Critical cultural awareness (*savoir s'engager*): the extent to which the student can evaluate critically and on the basis of explicit criteria perspectives, practices and products in one's own and other cultures, related to the foreign language literary text.

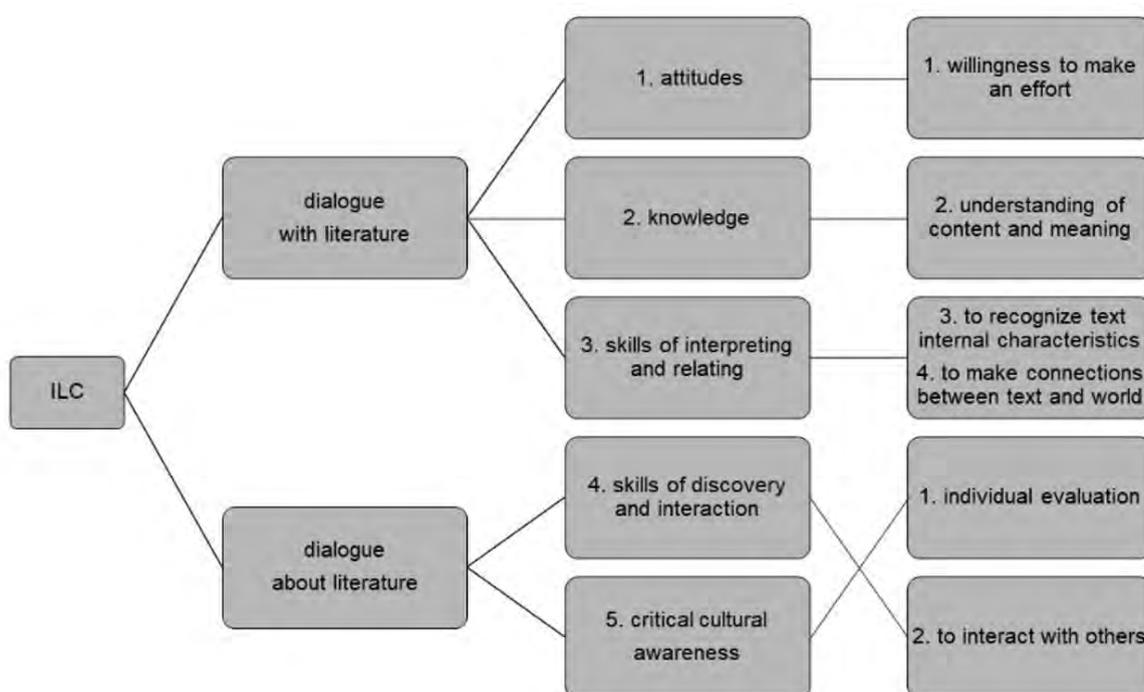**Figure 1** *Construct of intercultural literary competence*

Figure 2 *Schematic representation of the construct of intercultural literary competence*

It is noteworthy that foreign language proficiency is included in the presented construct and its accompanying learning objectives. While intercultural models often exclude target language

proficiency from the construct (Fantini, 2020), this model rests on the idea that content and language are learned in an integrated manner and that spoken and written L2 output about cultural content is essential to learn a foreign language (Schat *et al.*, 2020). Thus, in this specific conceptualization of intercultural competence being grounded in the domain of foreign language literature teaching, L2 proficiency is indispensable in all five ILC dimensions, as a good command of the target language facilitates both types of dialogue: students need appropriate reading proficiency in the target language to understand and interpret the literary text. Students need oral and writing proficiency to interact with peers or to express their views for example in written documents.

Current study and research questions

As can be seen from the explanation above, ILC is a new concept and involves a description of learning objectives specific for foreign language literature teaching at the secondary level. Our literature review showed that research focused mainly on higher education, and that studies aiming at the secondary level were either theoretical or lacking elaboration of learning objectives. Besides this theoretical gap, language teachers need evaluation tools to map the intercultural development of their students. In sum, an operational definition is needed. Following this rationale, we decided to develop a self-report questionnaire based on the ILC model, the Intercultural Literary Competence Questionnaire (ILCQ) and to validate this instrument. Despite the reliability and validity issues of self-reporting scales, as outlined in the theoretical framework above, we decided to construct a self-reporting tool. A clear advantage of a self-assessment is that it yields insights for both students and educators: the ILCQ can serve as a framework of reference for students and simultaneously provide teachers and researchers with a tool to diagnose intercultural progress or the effectiveness of educational practice. To our best knowledge, at the beginning of this study, no self-reporting instrument has ever been particularly developed to measure intercultural development through literature education for adolescents. Following this reasoning, two research questions were formulated:

1. What are the characteristics of an evaluation instrument that is valid as a measure of the construct of intercultural literary competence (ILC)?

2. How reliable is the Intercultural Literary Competence Questionnaire (ILCQ) used in the classroom context to map students' intercultural development?

Methodology

Item development

In order to operationalize the learning objectives for intercultural literary competence (ILC) into measurable items for a self-report questionnaire, we rewrote the student objectives in the form of can-do statements. As developing an adequate number of items is essential to encompass all aspects of a construct, we formulated five positive item statements for each ILC dimension. Thereupon, in order to ensure content validity, it is crucial to verify if the items in the preliminary pool actually cover the underlying theoretical framework. A common approach to examine content validity is to consult with experts. Accordingly, we asked three academics from the field of literature teaching or intercultural competence, and one curriculum developer (N=4) to revise the items. These experts were given the definition of the ILC dimensions and the list of 25 items and were asked to write their comments regarding how well the items covered the dimensions. Overall, the expert feedback was encouraging but raised issues on the dimensions of knowledge and skills of interaction. This led to the reformulation of item 7 and 18. As such, our initial Intercultural Literary Competence Questionnaire (ILCQ) consisted of 25 statements measured on a 7-point Likert scale. The response

options ranged from "completely disagree" to "completely agree." In each case, a higher score indicated a higher level of ILC. We chose a 7-point Likert scale, because this scale, as compared to a 5-point Likert scale, allows for more dispersion in the data, and more nuanced results can be obtained.

Another critical step in developing a new scale is to ensure that the destined respondents understand item wording and that can-do statements display clarity, providing for face validity. In order to pilot the 25 items, we set up think-aloud sessions (3 x 60 min) in an empty classroom with three volunteer students (two females and one male, aged 17 years) from the targeted audience. Think-aloud is a research method that supports researchers in the refinement and usability of instruments, as participants verbalize the thoughts that arise in their mind as they complete a task aloud (Van Someren *et al.*, 1994). One at a time, the students read aloud and responded to the items. During this process they were asked how they perceived and understood each item, regarding the wording and clarity of the can-do statements. At the end they were asked to provide an overall opinion of the list. Data collection for the think-aloud sessions included audio recording of the conversations and notes. The analysis of their feedback revealed that the students found the can-do statements comprehensible. Nevertheless, four items contained some subject-specific terminology (e.g. target language, evaluative analysis) which had to be adapted to more student-appropriate language. Regarding their overall opinion, they considered the 7-point Likert scale feasible but a 25-item list too long. Minor questionnaire changes were made to provide for face validity. Table 1 contains the 25 questionnaire items sorted according to the dimension to which they pertain. Subsequently, the preliminary ILCQ, as emerged from the abovementioned scale development process, was digitized. An online tool was developed using the program Formdesk. The language of the tool was Dutch and comprised three general questions (school, language and grade) and the 25 ILC items were presented in random order.

Participants and procedure

As the aim of this research is twofold, validating an instrument and investigating its reliability for classroom use, this study was set up in two phases: a development phase and an implementation phase. In the development phase it was investigated to which extent the ILCQ represented the construct of ILC, while the implementation phase investigated the reliability of the ILCQ. In order to address both research questions, data collection comprised two questionnaire rounds with a different sample for each administration. In the development phase, in which factor analysis was carried out to determine the construct validity of the preliminary instrument, the 25-item ILCQ was administered anonymously to foreign language students (N=164) in various secondary schools. In the implementation phase, which focused on the prospective use of the revised ILCQ as a pre- and post-test for classroom use, data collection took place in a class size sample (N=20) and was not anonymous in order to perform test and retest analysis.

For the first phase of data collection, which was set in January 2019, an invitation to participate was sent to foreign language teachers using contacts in the first author's network. Teachers were asked to distribute the ILCQ among their students and they could opt to either have the questionnaire completed in the classroom or distribute the link online. The ILCQ was administered to 203 students from six secondary schools. After eliminating incomplete and poorly responded questionnaires, the sample for the first phase consisted of 164 secondary school students. Participants (aged 16-19) were drawn from tenth grade (25%), eleventh grade (41%) and twelfth grade (34%). The second languages addressed in the questionnaire were Spanish (73%), followed by German (15%) and French (12%). Participating schools were spread all over the country in both rural and urban areas (3 in the west region, 1 in the east, 2 in the south).

Table 1 *The 25 questionnaire items for intercultural literary competence*

Labels	Questionnaire items
a_1	I am interested in daily life in other cultures as depicted in the book.
a_2	I am willing to empathize with characters in the book.
a_3	I find it interesting to discover other perspectives on interpretation of familiar and unfamiliar phenomena in the book.
a_4	I am interested to discover in the book how people from other cultures would view practices and products of my culture.
a_5	I am willing to take up another perspective and critically examine my own cultural perspective on the book.
k_1	I can use the book to learn more about the relationship between my culture and other cultures.
k_2	I can use the book to learn more about stereotypes and prejudices concerning my culture and other cultures.
k_3	I can use the book to learn more about different perspectives on national memory and national definitions of geographical space.
k_4	I can use the book to learn more about different social groups and society in other cultures.
k_5	I can use the book to learn more about conventions of behavior and individual interaction in other cultures.
ip_1	I can describe important characters/events in the book and identify areas of cultural misunderstanding.
ip_2	I can elaborate on the theme of the book and explain how different cultural positions make different interpretations.
ip_3	I can explain the effects of narrative perspective and identify ethnocentric perspectives in the book.
ip_4	I can relate the characters/events/themes of the book to my own life despite cultural differences.
ip_5	I can relate the themes in the book to current social issues and explain sources of cultural misunderstanding.
ia_1	I can discuss with my fellow student our interpretations establishing relationships of similarity and difference between them.
ia_2	I can prepare and carry out a dialogue with my fellow student in which I can take up the perspective of a character from the book.
ia_3	I can ask my fellow student about shared meaning and values relating to the book and establish links and relationships among them.
ia_4	I can discuss with my fellow student our opinions on the book and establish relationships of similarity and difference between them.
ia_5	I can use other sources of information to learn more about the book and its context, and analyze different interpretations involved.
ca_1	I can write an evaluative analysis on the book, placing the book in context and demonstrating the ideology involved.
ca_2	I can write a personal reaction to the book, making a judgement with explicit reference to my own ideological perspectives and values.
ca_3	I can mediate different evaluations of the book, negotiating agreement on places of conflict and acceptance of differences.
ca_4	I can evaluate the book with respect to how the book challenged my cultural assumptions and to which personal insights it has led.
ca_5	I can judge how the book makes it possible for me to identify with characters from other cultures and give examples.

Note: The five subdimensions of ILC are labeled as attitude (a), knowledge (k), interpretation (ip), interaction (ia) and cultural awareness (ca).

In the second phase of data collection, in which we wanted to investigate whether our revised version of the ILCQ was a stable measure of the intercultural development of secondary school students in a classroom setting, the sample comprised 20 students from two Spanish-as-a-Foreign-Language classes (tenth grade, N=12 and eleventh grade, N=8) at two schools. The first test was set at the end

of January and the retest was taken in the beginning of March 2019. All participants filled out the form during regular class time. Of the sample, 7 were male and 13 were female students. Although participation was voluntary and participants were guaranteed confidentiality, data collection in this phase was not completely anonymous as identification was indispensable for test and retest analysis. Students were asked to give their written consent and informed that all data were used only for research purposes. The study was approved by the Ethical Commission of our University.

Data analysis

Validity in the development phase

In order to investigate the validity of a questionnaire, factor analysis is recommended as a method (Costello & Osborne, 2005). When developing a new scale, it is advisable to first employ exploratory factor analysis (EFA) before performing confirmatory factor analysis (CFA), as EFA generates hypotheses exploring the factor structure, whereas CFA tests hypotheses and corroborates the existence of a relation between the observed variables and the drawn models. However, EFA can also be used as a data reduction technique ensuring that the most important and representative items are selected. Considering that the first aim of this study was to test if the ILCQ represented the theoretical model of ILC, the purpose of EFA in this study was not to generate hypotheses on the factor structure but rather to refine the item pool, making a good pre-selection of representative items of ILC for the subsequent CFA Shorter instruments have obvious benefits like rapid implementation and ease of interpretation, leading to reduced student and teacher burden. Given that we wanted to make the ILCQ a feasible tool for classroom use and that a 25-item list was considered long by the students in the think aloud sessions, EFA was merely used to discard items, minimizing the duration of the survey. Any item with a factor loading less than .32 or with high factor loadings on more than one factor were discarded from the preliminary item pool (Costello & Osborne, 2005). For the EFA, a Principal Axis Factoring (PAF) with oblique (oblimin) rotations with Kaiser normalization was performed using SPSS software.

To investigate the construct validity of the questionnaire we performed a CFA with AMOS software. CFA allows us to test the hypothesis that a relationship between observed variables and their underlying latent constructs exists. As our purpose was to test if our instrument represented the hypothesized model of ILC with two categories (dialogue with literature and dialogue about literature) and its five subdimensions, we tested a two second-order model with five first order factors. We also tested two alternative models to assess alternative hypotheses for the ILCQ. The first alternative model was a one factor latent model with all items. The second alternative was a five factor first-order model based in the five subdimensions. For the CFA, we employed the following goodness-of-fit indices to assess model fit: Relative chi-square (CMIN), the Comparative Fit Index (CFI), the Tucker Lewis Index (TLI), the Root Mean-Square Error of Approximation (RMSEA) and Standardized Root Mean Residual (SRMR). The CMIN statistic should be below 5, but is ideally below 2. The CFI and the TLI should be above .90 and the RMSEA and SRMR should be below .08 (Marsh *et al*., 2004). Hu and Bentler (1999) propose more stringent cut-off values (CFI/TLI $\geq$ .95, RMSEA $\leq$ .06 and SRMR $\leq$ .08) when developing a scale.

Reliability analysis in the implementation phase

In the implementation phase, the reliability of the revised ILCQ was tested using a test-retest design with a 4-week interval between measurements. We calculated three indices: Cronbach's alpha (α), Pearson's correlation coefficient (*r*), and summability (*s*). The first measure of reliability, Cronbach's alpha, is a measure for the internal consistency of the subscales. Another form of reliability is

stability, or test-retest reliability, which was assessed by examining the correlation coefficient between the scores of the first and second measurement in the implementation phase. The third index, summability, is a measure of the quality of the sum score as a summary of the test (Goeman & de Jong, 2018).

Results

Validity in the development phase

Item reduction with exploratory factor analysis

Prior to performing EFA, the suitability of the respondent data in the development sample was assessed by performing two prerequisite tests for factor analysis: The Kayser-Meyer-Olkin (KMO) and Bartlett's sphericity test. The KMO test measures the sampling adequacy. For a factor analysis, the KMO must be at least .50; KMO values above .80 are ideal (Williams *et al.*, 2010). Bartlett's sphericity test checks for a redundancy between variables that can be summarized by other factors. The KMO value was .95 and Bartlett's sphericity test was significant ($X^2 = 2961.95$, $df = 300$, p < .001). In other words, the dataset revealed a good sample size and sufficient correlation between the variables to continue with the EFA. In addition, we calculated the Cronbach's alphas of the total scale and for each of the five dimensions to explore whether our items could be considered to form subscales. The total scale ($\alpha = .96$) and each of the five subscales ($\alpha = .85$ to $\alpha = .89$) demonstrated high levels of internal consistency.

Table 2 *Results of exploratory factor analysis*

Labels	Factors				
	1	2	3	4	5
ip_1	**0.49**	-0.12	0.08	-0.17	0.21
ip_2	**0.34**	-0.22	0.25	-0.02	0.10
ip_5	**0.34**	-0.03	0.22	-0.21	0.15
a_3	-0.02	**-0.93**	0.03	0.08	0.05
a_4	-0.01	**-0.68**	0.07	-0.13	-0.12
a_1	0.11	**-0.62**	-0.08	-0.05	0.09
a_2	-0.14	**-0.35**	0.24	-0.18	0.29
ia_4	0.02	-0.04	**0.84**	-0.09	-0.14
ia_1	0.02	0.00	**0.73**	0.07	0.20
ia_2	0.27	-0.05	**0.40**	0.04	0.29
k_4	-0.01	-0.01	0.03	**-0.82**	0.12
k_5	0.21	-0.30	0.00	**-0.50**	-0.04
ca_4	0.10	-0.01	0.00	-0.16	**0.70**
ca_1	0.07	-0.07	0.06	-0.07	**0.60**
Eigenvalues	7.60	1.27	0.95	0.57	0.54
%variance	54.33	9.05	6.75	4.01	3.88
α (Cronbach's alpha)	.82	.85	.82	.84	.79

As the objective of the EFA was not to explore factor structure, but to select the most representative items for each of the five subdimensions of ILC, we manipulated the number of factors to extract at five, hereby disregarding the first results of the scree-test and the Kaiser-criterium, which suggested the extraction of three factors. Literature on the topic (Costello & Osborne, 2005) suggests setting the number of factors to manually retain and running multiple analyses when the number of factors

suggested by the scree-test is different from the predicted a priori factor structure. Subsequently, items that had either low factor loads (< .32) or high loadings on more than one factor were excluded from the scale. Firstly, item ia_5 with a value below the cut-off point was deleted from the scale. Subsequently, cross loading items were discarded one by one, starting with the most cross-loading one. In the process of all these analyses a total of 11 items was deleted from the preliminary scale: a_5, k_1, k_2, ip_3, ip_4, k_3, ca_3, ca_5, ca_2, ia_3 and ia_5.

Table 2 summarizes the results of the EFA with PAF showing the remaining 14 items with their factor loadings (.34 - .93). As our aim was to develop a short questionnaire in which the five subdimensions of ILC were covered, we found a workable solution with 14 items that loaded on five distinct factors. Although the solution revealed only two factors with eigenvalues over 1, we decided to continue with this item selection for the CFA as the purpose of our EFA was not to explore factorial structure, but rather to select the most representative items. As can be seen the total variance explained by the five-factor solution with 14 items was 78% and the remaining dataset maintained sufficient correlation (X^2 = 1417.464, *df* = 91, p < .001) and sample size (KMO = .94) for CFA. In accordance with the adopted framework of ILC, we labelled the five factors attitude (a, 4 items), knowledge (k, 2 items), interpretation (ip, 3 items), interaction (ia, 3 items) and cultural awareness (ca, 2 items).

Conforming factorial structure with confirmatory factor analysis

With this selection of items, we ran a CFA to explore how these variables relate to the construct of ILC. As suggested by Costello and Osborne (2005), a strong factor demands a minimum of three items; thus, prior to the CFA one item was added to the subscales of knowledge (k_1) and cultural awareness (ca_2), resulting in a total of 16 items for the revised ILCQ. The 16 selected items are highlighted in bold font in Table 1. Three models were examined to investigate which structure best represents the construct of ILC. First, a unidimensional one factor model in which all items loaded on one latent factor. Second, a multidimensional five factor model in which the items loaded on five first-order latent factors, the five subdimensions of ILC. And third, the a-priori hypothesized model, which is a hierarchical model composed of two second-order factors (the categories dialogue with literature and about literature) and five first-order factors (the five subdimensions).

Table 3 *Goodness of fit indices*

Model	Fit Indices						
	X^2	*df*	CMIN	CFI	TLI	RSMEA	SRMR
model 1 (one factor)	596.27	106	5.63	.71	.67	.17	.27
model 2 (five factor)	160.95	94	1.71	.96	.95	.06	.05
model 3 (two second-order)	193.19	98	1.97	.94	.93	.08	.05

CMIN = Relative chi-square; CFI = Comparative Fit Index; TLI = Tucker Lewis Index; RMSEA = Root Mean-Square Error of Approximation; SRMR = Standardized Root Mean Residual

As can be seen in Table 3, the one factor model, with factor loadings ranging from .63 to .78, showed a very poor fit with the data (X^2 = 596.27, *df* = 106, p < .001, CFI = .71, TLI = .67, RMSEA = .17, SRMR = .27) indicating that the ILCQ measured a multidimensional construct. Model 2 and 3 do meet the criteria of adequate fit (Marsh *et al*., 2004) as they have CMIN below 2, CFI above .90, TLI close to 1.0., RMSEA and SRMR below .08. Both drawn models fit well on the data: all loadings are above .71 and are significant (> .70). Nevertheless, model 2 (X^2 = 160.95, *df* = 94, p < .001, CFI =

.96, TLI = .95, RMSEA = .06, SRMR = .05) has a better fit than model 3 (X^2 = 193.19, *df* = 98, p < .001, CFI = .94, TLI = .93, RMSEA = .08, SRMR = .05). If we use the more stringent cut-off values (CFI/TLI ≥ .95, RMSEA ≤ .06 and SRMR ≤ .08) proposed by Hu and Bentler (1999), only the five factor structure had a good fit.

Table 4 *Results of implementation phase*

Labels	Measurement 1			Measurement 2					
	Mean	SD	α	Mean	SD	α	*r*	*s*	range
attitude_scale	4.58	1.09	.76	4.61	1.13	.79	.69	.50	1-7
a_1	5.15	1.27		4.80	1.51				
a_2	4.05	1.10		4.00	1.59				
a_3	4.40	1.76		4.85	1.23				
a_4	4.70	1.53		4.80	1.44				
knowledge_scale	4.13	1.29	.80	4.00	1.32	.74	.71	.59	1-7
k_1	4.15	1.39		4.00	1.38				
k_4	3.90	1.29		4.15	1.63				
k_5	4.35	1.84		3.85	1.46				
interpretation_scale	3.87	1.21	.70	4.20	1.14	.78	.43	.55	1-7
ip_1	4.05	1.70		4.70	1.49				
ip_2	3.75	1.45		4.00	1.21				
ip_5	3.80	1.44		3.90	1.41				
interaction_scale	4.12	1.40	.77	4.12	1.60	.88	.72	.72	1-7
ia_1	4.00	1.75		4.05	1.73				
ia_2	3.95	1.67		4.15	1.93				
ia_4	4.40	1.67		4.15	1.66				
cultural_awareness_scale	3.95	1.14	.69	4.48	1.07	.83	.59	.62	1-7
ca_1	4.40	1.27		4.45	1.36				
ca_2	3.50	1.36		4.40	1.14				
ca_4	3.95	1.67		4.60	1.35				
Total	4.15	1.00	.91	4.27	1.08	.93	.74	.50	1-7

Thus, according to the fit indices, a first-order five factor model with CFI and TLI above .95 and a RSMEA of .06 was considered to be the best explanation of the sample data while the hierarchical two second-order model also had acceptable fit with CFI and TLI above .90 and RSMEA below .08. These findings reveal that a five factor model as well as higher order model are usable to measure ILC. However, in both models the inter-factor correlations were high. In accordance with the ILC theoretical framework, all factors were allowed to correlate with each other. In the five factor model inter-factor correlation ranged from .68 to .96 and in the two second-order model the correlation between dialogue with and about literature was .96. Due to this very high conceptual overlap between the five first-order factors and the two second-order factors, questions may be raised regarding the discriminant validity of each of the factors. This means that scores on the five the subscales of the ILCQ can be safely used for descriptive statistics or as pre- and post-test measures, yet not to be used as predictor variables due to issues of multicollinearity. Multicollinearity is a phenomenon in which one independent variable is highly correlated with one or more of the other independent variables. Given the fact that the ILCQ scores will predominantly be used as a measure of learning outcomes, that is to say as dependent variables, multicollinearity does not pose a threat to the model.

Reliability analysis in the implementation phase

After the validity of the 16-item questionnaire was investigated, the instrument was set out twice in two groups to investigate consistency, stability and summability. Table 4 shows the resulting ILCQ with mean values, standard deviations, alphas, correlation coefficients, summability measures and range. The first measurement showed that the internal consistency of the total scale was good (α = .95). Also, all ILCQ subscales showed coefficients that were nearly sufficient to relatively high (.69 to .80) as evidence of internal consistency. Also at the second data collection point, all scales were internally consistent and well-defined by their items, as indicated by Cronbach's alphas well above the threshold value of .70 (α = .74 to α = .93) Furthermore, the correlation coefficient for the total scale was highly significant (r = .74), as well as for the subscales (r = .59 to r =.72), except for the interpretation scale (r = .43). From the high test-retest correlations we deduced that using the ILCQ as a pre-test would yield an acceptable prediction for the post-test. From the summability indices (s = .50 to s = .72) we concluded that all items in the ILCQ measure ILC. In addition, the width of the range can also be seen as an indicator of a reliable instrument. As such, the final draft of the ILCQ, with 16 items divided into five domains, demonstrated optimum psychometric properties.

Discussion

Conclusions

The aim of this study was to develop and validate a self-reporting intercultural competence evaluation instrument for foreign language literature teaching. In order to do so, the Intercultural Literary Competence Questionnaire (ILCQ) was developed based on the construct of intercultural literary competence (ILC). The theoretical model of ILC implies a two-level hierarchical structure in which the five *savoirs* of intercultural communicative competence (Byram, 1997) form the five first-order factors; the distinction between communicating with literature and communicating about literature (Witte, 2008) form the two second-order factors. To investigate the validity of the instrument, the current study tested this structure as well as two alternative structures, a five factor first-order model and a unidimensional model. Based on the theory of ILC, we expected that the hierarchical model would provide the best fit for the data. In a later phase of the study, the instrument was set out in a small sample to perform reliability analysis.

Results from the first phase indicated that the a priori two second-order factor structure of the ILCQ showed acceptable model fit. Nevertheless, the model fit indices indicated that ILC was better represented by a five-factor model with the latent variables attitude, knowledge, interpretation, interaction and cultural awareness. This model represents the structure of intercultural communicative competence as described by Byram (1997) that labelled five dimensions for the construct. In this five factor first-order model the standardized factor loadings for all items were high and statistically significant. This means that all items are a good reflection of the factors, which points to excellent convergent validity. Nevertheless, the fact that the five factors were all highly correlated points to conceptual overlap of the five dimensions and may raise issues regarding the discriminant validity of the ILCQ subscales. Accordingly, while measures of the five first order factors and the two second order factors may be used for descriptive statistics or as pre- and post-test measures, they may show no significant predictive utility above and beyond that of a general higher-order ILC factor in terms of their association with other outcome variables. The results of the second phase corroborated the results of the first phase.

Considering the high correlations between the five dimensions found in the development phase and the good summability indices found in the implementation phase, the overall findings of this study point to the appropriateness of this tool as a single measurement to express a student's ILC through

the sum score of the total scale. This is highly convenient for teachers in the context of a classroom to give a quick indication of intercultural development. Teachers can administer the digital scale easily and generate one variable rather than producing five variables with subscale scores. The use of the ILCQ for the purpose of obtaining an overall measure of a student's ILC is thus recommended. Nevertheless, the goodness of fit indices of the five factor first-order model, added to the high consistency and test-retest reliability of the subscales, reveal that the ILCQ can also be used as a measure for the five dimensions of intercultural competence. While some scholars (Hoff, 2020; Liddicoat & Scarino, 2013) argue that Byram's model of intercultural competence lacks clarity regarding the extent to which the dimensions are integrated and that, for that matter, operationalization for assessment remains an issue, the results of this study indicate the feasibility of the five-dimension model for assessment. In addition, it shows that CFA is an accurate method for empirically testing different intercultural models, as it "can help to identify the structure of multicomponent measures of intercultural competence" (Van de Vijver & Leung, 2009, p. 415). Although the more complex two second-order model—with the addition of dialogue with literature and about literature—did not increase model fit, this distinction might also be useful for the literature classroom. As some students may have more difficulties in interpreting a book or others might find it more difficult to share their opinions with others, teachers may use these scores to access student level or as a diagnostic tool for implementing appropriate curricular activities in the literature classroom. Therefore, the choice of which model of the scale to be used depends on its purpose.

Limitations and implications for future research

As can be concluded from above, the practical implications of this study for classroom purposes are substantial. Since the digital ILCQ tool was designed and validated on the target population, it can directly be applied in foreign language classrooms at the secondary level. Although the scale that we developed has arisen from the specific context of the Netherlands and was intended to serve the needs of Dutch foreign language teachers in the first place, the ILCQ has potential widespread application on an international level given how commonly foreign languages are taught in secondary schools around the world. Translation and testing in other languages is therefore suggested for further validation. Furthermore, a clear score to estimate the degree of ILC not only benefits teaching practice but can also have strong implications for research. As the current study has focused on operationalizing the contribution of literature education to intercultural development, researchers can use the instrument as a pre-and post-test for literature classroom intervention studies and support their theories with statistical evidence. In such a way, the quality of classroom studies can be improved (Schrijvers *et al*., 2019). There is now a sizable body of research on literature and interculturality, but no instruments for assessing intercultural competence as an outcome of literature programs. This study and instrument can enhance the current discourse about the benefits of literature for intercultural development (Matos & Melo-Pfeifer, 2020; Porto & Zembylas, 2020) with empirical arguments. Besides providing a sound assessment instrument, this study has also yielded a theoretical model which may inform intercultural theory or serve as a model of reference for literary pedagogy.

Whereas this study has demonstrated the validity, reliability and utility of the instrument, a number of limitations regarding the generalizability of these results should be noted. The first limitation concerns the method of factor analysis. As it is a sample-dependent technique, caution is needed when interpreting the results. The ILCQ was reduced from 25 to 16 items based on the data of this specific sample. Several items were removed from the ILCQ because they loaded onto more than one factor. More research is needed to explore whether this reduction was specific to our sample. In addition, factor analysis is heuristic as it relies heavily on judgements made by researchers (Williams *et al*., 2010). While our purpose was to verify whether the ILCQ represented the five dimensions of

ILC, studies with distinct purposes might find other latent variables. Interpreting what the sets of variables actually represent is up to the researcher as more than one interpretation can be made of the same data factored the same way.

A second shortcoming concerns the sample. As the teachers in this study decided voluntarily to participate in this study and chose to distribute the ILCQ among their students, the sample of students used is a convenience sample. For this reason, one might argue that evidence that the instrument is applicable to foreign language students originated with teachers who are interested in intercultural competence and literature. However, as the model aims exactly at this population, this should not pose a threat to the validity of the model. As the practical aim of this study was to design a feasible instrument for foreign language teachers to monitor their students' intercultural development through the reading of foreign language literary texts, the ILCQ can be considered a valid and reliable instrument: it measures what it is supposed to measure and produces results that can be trusted in the target context. Nevertheless, whereas the results of this study point to the viability of the ILCQ and argue for its use in the Dutch secondary education context, the ILCQ would benefit from further validation with different samples in different educational and cultural contexts. Various studies have highlighted the poor generalizability of intercultural competence instruments, as they fail to account for different cultural contexts (Deardorff, 2016; Perry & Southwell, 2011). As the validity of the ILCQ in other contexts cannot be taken for granted, we suggest translating the ILCQ into other languages and investigating its utility in various cultural contexts.

A final limitation is inherent to the construct of intercultural competence itself and concerns validity. Considering the most common definition of intercultural competence as "the ability to behave and communicate effectively and appropriately" (Deardorff, 2016, p. 122), self-reports only measure half of the picture as they cannot assess appropriateness. Therefore, it is suggested that the ILCQ be triangulated with other forms of assessment. The use of multiple combinations of assessment types, such as self-assessment with formative teacher-led assessment, or qualitative and quantitative evaluation is therefore strongly recommended (Liddicoat & Scarino, 2013). Researchers should not only validate this instrument in different samples or other educational contexts, but also complement it with other types of ILC assessment. We hope that the theoretical model of ILC serves as a model or reference for both teachers and researchers to develop different forms of assessment. Future research might focus on how students achieve ILC objectives through a combination of lesson observations or text analysis of student assignments.

In conclusion, our study produced a reliable and psychometrically valid scale for evaluating ILC. The ILCQ, to the best of our knowledge, is the first intercultural competence instrument specifically designed for literature classrooms at the secondary level. Aside from being a viable tool for teachers and researchers, the use of ILCQ can promote adolescents' awareness of the possibilities that literature can offer for nurturing empathic abilities and cultivating tolerance. Discussing the relevance of the ILC statements in the classroom can stimulate the debate about why literature should be an integral part of the curriculum and why literature is so urgently important for the citizen (Nussbaum, 1998). In this way, the development of this instrument may have a considerable backwash effect, reinforcing the place of intercultural competence on the language-teaching-agenda and gaining broader support for the study of literary texts in secondary education.

Acknowledgements

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Alter, G., & Ratheiser, U. (2019). A new model of literary competences and the revised CEFR descriptors. *ELT Journal*, *73*(4), 377–386. https://doi-org.proxy.library.uu.nl/10.1093/elt/ccz024

Barrett, M., Byram, M., Lázár, I., Mompoint‑Gaillard, P., & Philippou, S. (2014). *Developing intercultural competence through education*. Strasbourg: Council of Europe.

Biesta, G. J. J. (2006). *Beyond learning. Democratic education for a human future.* Boulder: Paradigm.

Bloemert, J., Paran, A., Jansen, E., & van de Grift, W. (2019). Students' perspective on the benefits of EFL literature education. *The Language Learning Journal*, *47*(3), 371–384. https://doi.org/10.1080/09571736.2017.1298149

Bredella, L. (2008). What makes reading literary texts pleasurable and educationally significant? *FLuL–Fremdsprachen Lehren und Lernen*, *37*, 12–26.

Burwitz-Melzer, E. (2001). Teaching intercultural communicative competence through literature. In M. Byram, A. Nichols, D. Stevens (Eds.), *Developing intercultural competence in practice* (pp. 29–43). Clevedon: Multilingual Matters.

Byram, M. (1997). *Teaching and assessing intercultural communicative competence*. Clevedon: Multilingual Matters.

Byram, M., & Wagner, M. (2018). Making a difference: Language teaching for intercultural international dialogue. *Foreign Language Annals*, *51*(1), 140–151. https://doi.org/10.1111/flan.12319

Costello, A., & Osborne, J. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment Research & Evaluation*, *10*(7), 1–7. https://doi.org/10.7275/jyj1-4868

Council of Europe. (2018, February). *Common European Framework of Reference for Languages: learning, teaching, assessment–companion volume with new descriptors.* Retrieved from https://rm. coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Curriculum.nu. (2019, October). *Vierde tussenproduct ontwikkelteam Engels/MVT Curriculum.nu.* Retrieved from https://www.curriculum.nu/voorstellen/engels-mvt/

Deardorff, D. K. (2016). How to assess intercultural competence. In Zhu Hua (Ed.), *Research methods in intercultural communication. A practical guide* (pp. 120–134). London: Wiley Blackwell.

Delanoy, W. (2005). A dialogic model for literature teaching. *ABAC Journal, 25*(1), 53–66.

Fantini, A. E. (2020). Language: an essential component of intercultural communicative competence. In J. Jackson (Ed), *The Routledge handbook of language and intercultural communication*, (pp 263–278). London & New York: Routledge.

Goeman, J. J., & de Jong, N. H. (2018). How well does the sum score summarize the test? Summability as a measure of internal consistency. *Educational Measurement: Issues and Practice*, *37*(2), 54–63.

Hoff, H. E. (2016). From "intercultural speaker" to "intercultural reader": A proposal to reconceptualize intercultural communicative competence through a focus on literary reading. In *Intercultural competence in education* (pp. 51–71). London: Palgrave Macmillan. https://doi.org/10.1057/978-1-137-58733-6_4

Hoff, H. E. (2020). The evolution of intercultural communicative competence: Conceptualisations, critiques and consequences for 21st century classroom practice. *Intercultural Communication*

Education, *3*(2), 55–74. https://doi.org/10.29140/ice.v3n2.264

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modelling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Kramsch, C. (1993). *Context and culture in language teaching.* Oxford: Oxford University Press.

Lehrner-te Lindert, E. (2020). Fremdsprachliches Lesen mit literarischen Texten. Zur Entwicklung von Leseverstehen und literarischer Kompetenz im DaF-Unterricht der niederländischen Sekundarstufe I. Berlin: Erich Schmidt Verlag.

Liddicoat, A. J., & Scarino, A. (2013). *Intercultural language teaching and learning.* Chisester: John Wiley and Sons.

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modelling: A Multidisciplinary Journal*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Matos, A.G., & Melo-Pfeifer, S. (2020). Intercultural language education for the 21st century through literary texts. In A.G. Matos & S. Melo-Pfeifer (Eds.), *Literature and intercultural learning in language and teacher education* (pp. 9–20). Bern: Peter Lang.

Meesterschapsteam MVT. (2018, March). *Visie op de toekomst van het curriculum Moderne Vreemde Talen.* Retrieved from https://modernevreemdetalen.vakdidactiekgw.nl/wpcontent/uploads/sites/6/2018/03/Aanbevelingstekst-MT-MVT.pdf

Meijer, D., & Fasoglio, D. (2007). *Handreiking schoolexamen moderne vreemde talen havo/vwo.* Enschede: Netherlands Institute for Curriculum Development (SLO).

Nussbaum, M. C. (1998). *Cultivating humanity. A classical defense of reform in liberal education.* Harvard: Harvard University Press.

Perry, L. & Southwell, L. (2011) Developing intercultural understanding and skills: models and approaches. *Intercultural Education*, *22*(6), 453–466. https://doi.org/10.1080/14675986.2011.644948

Porto, M., & Zembylas, M. (2020). Pedagogies of discomfort in foreign language education: cultivating empathy and solidarity using art and literature. *Language and Intercultural Communication*, *20*(4), 356–374. https://doi.org/10.1080/14708477.2020.1740244

Risager, K. (2007). *Language and culture pedagogy: From a national to a transnational paradigm* (Vol. 14). Clevedon: Multilingual Matters.

Rodríguez, L. M. G., & Puyal, M. B. (2012). Promoting intercultural competence through literature in CLIL contexts. *Atlantis*, *34*(2), 105–124.

Schat, E., van der Knaap, E., & de Graaff, R. (2020). Literatuuronderwijs ter bevordering van interculturele competentie; een pilotonderzoek. *Levende Talen Tijdschrift*, *21*(2), 13–25.

Schat, E., de Graaff, R., & van der Knaap, E. (2018). Intercultureel en taalgericht literatuuronderwijs bij de moderne vreemde talen; Een enquête onder docenten Duits, Frans en Spaans in Nederland. Levende Talen Tijdschrift, *19*(3), 13–25.

Schrijvers, M., Janssen, T., Fialho, O., & Rijlaarsdam, G. (2019). Gaining insight into human nature: A review of literature classroom intervention studies. *Review of Educational Research*, *89*(1), 3–45. https://doi.org/10.3102/0034654318812914

Sercu, L. (2004). Assessing intercultural competence: A framework for systematic test development in FL education and beyond. *Intercultural Education*, *15*(1), 73–89. https://doi.org/10.1080/1467598042000190004

Spitzberg, B. H., & Changnon, G. (2009). Conceptualizing intercultural competence. In D. Deardorff (Ed.), *The SAGE handbook of intercultural competence* (pp. 2–52). Los Angeles: Sage.

Van Someren, M., Barnard, Y., & Sandberg, J. (1994). *The think aloud method: a practical approach to modelling cognitive.* London: Academic Press.

Van de Vijver, F., & Leung, K. (2009). Methodological issues and researching intercultural

competence. In D. Deardorff (Ed.), *The SAGE handbook of intercultural competence* (pp. 404–417). Los Angeles: Sage.

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. *Australasian Journal of Paramedicine*, *8*(3), 1–13. https://doi.org/10.33151/ajp.8.3.93

Witte, T. C. H. (2008). *Het oog van de meester: een onderzoek naar de literaire ontwikkeling van HAVO-en VWO-leerlingen in de tweede fase van het voortgezet onderwijs*. Delft: Eburon.

Author biodata

Esther Schat works as a Spanish teacher in secondary education and is currently engaged in doctoral research on the topic of interculturality and literary pedagogy at the faculty of Humanities of Utrecht University in a scholarship program for teachers from Dudoc Alfa. She holds master's degrees in foreign language education and Latin American studies.

Ewout van der Knaap is a professor in German Literature and Culture at Utrecht University. He conducts research in the field of literary pedagogy.

Rick de Graaff is a professor of foreign language pedagogy and multilingual education at Utrecht University and the University of Applied Sciences Utrecht. His research focuses on language pedagogy, content and language integrated learning and the professional development of language teachers.