CHAPTER 10

Guidelines for the Validation of Writing Assessment in Intervention Studies

Renske Bouwer, Elke Van Steendam and Marije Lesterhuis

To draw valid conclusions about the effectiveness of writing interventions, the assessment design is key. Within a writing assessment, observations of students' performance on a particular writing task are generally used to make inferences about their writing proficiency. To do that, researchers have to make various decisions, e.g., about the kind of text(s) students have to write, how the texts are scored and by whom. Moreover, they have to underpin these decisions with arguments and evidence to make sure the conclusions they draw based upon text quality scores are valid interpretations of students' writing proficiency. This requires a thorough and comprehensive analysis of theoretical, empirical, and analytical evidence of writing assessment information, which is often not manageable for researchers whose primary focus is on the writing intervention (Shaw et al., 2012).

The aim of this chapter is to provide specific guidelines for researchers on how to make evidence-based choices when assessing writing in a writing intervention study. As such, the guidelines are meant to empower researchers to argue and support the validity of their decisions in and about the assessment of writing. To do that, we will use an existing framework for an argument-based approach to validity (Kane, 2006; 2011) and apply it to writing by reviewing theoretical and empirical evidence from writing assessment research.

The chapter starts with a brief discussion of Kane's framework that presents a chain of inferences which researchers need to make on the basis of assessment scores. An in-depth discussion of Kane's framework remains outside this chapter's scope, but we will use it to specify the inferences that are generally made within the context of a writing intervention study and the kind of evidence that is needed to substantiate those inferences. For each proposed inference, we will use insights from writing assessment research to provide evidence-based guidelines for researchers to carefully select the most appropriate assessment procedures and to build their own case for the validity of these inferences.

1 Applying Kane's Validity Framework to Writing Assessment

Key to Kane's framework, and crucial for this chapter, is that validity goes beyond the question which rating procedure, what task and how many raters to select. Validity is not a characteristic of an assessment or an assessment procedure, but it is about the extent to which assessment scores are appropriately used and interpreted. Kane (2006, 2011) states that validity needs to be argued, that is, evidence-based arguments need to be provided to validate the claims about the effectiveness of a specific intervention on the basis of assessment scores. What makes such a line of reasoning particularly challenging is that the purpose or claims of the intervention may differ in each study. In other words, researchers have to determine for each intervention study what conclusions they want to draw on the assessment scores, and then have to provide arguments and underlying evidence that support the validity of these conclusions.

In order to underpin such validity claims in a systematic manner, Kane's framework presents a stepwise chain of reasoning along five inferences. These inferences can assist in deciding which evidence should be acquired to build a sound argumentation for the validity of score interpretation and use. The first step in this line of reasoning is to justify whether inferences from the domain of interest to observations within the assessment are warranted. This inference is based on the assumption that the tasks that are used to observe or measure students' performance are an adequate reflection of all possible tasks within the domain of interest (i.e., task inference). Evidence is needed to provide support for this first assumption. In subsequent steps, it should be argued whether scores for the observed performance can be interpreted as intended (i.e., scoring inference), can be generalized to the broader test domain (i.e., generalization inference), can be extrapolated to the more general practice domain (i.e., extrapolation inference), and finally, can be used to make the intended decisions (i.e., decision inference). Together, these inferences form a chain of argumentation in which each of the inferences extends the use or interpretation of scores. As a result, "the overall argument is only as strong as its weakest link" (Kane et al., 1999, p. 15). Figure 10.1 illustrates how the chain of reasoning along five inferences can be applied to writing assessment in intervention studies.

Below, we will provide specific guidelines for justifying the plausibility and appropriateness of these five inferences by reviewing available theoretical and empirical evidence in previous writing assessment research. It is important to note that inferences can vary from study to study, depending on how assessment scores are interpreted and used. Because of this variety, it is impossible to provide general rules for writing assessments that guarantee valid interpre-

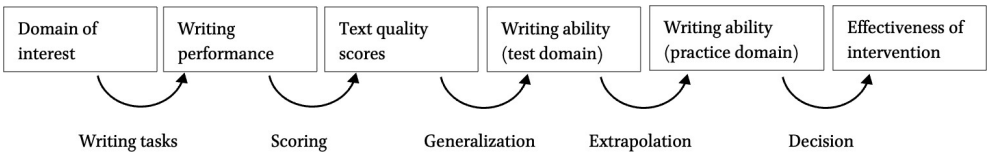

FIGURE 10.1 General validation framework of writing assessment in an intervention study

tations in every intervention study. This implies that researchers have to specify which of the proposed inferences are made in their intervention study, and to build their own case for the validity of the assumptions within those inferences.

Also, in some writing studies, the assessment is part of the intervention, for instance when assessment scores are used formatively by teachers to adapt their writing instructions or to provide feedback to students (cf. chapter 9 by Arrimada, 2023 in this volume). This involves additional steps in the validity framework, which also require additional validation evidence, such as whether teachers make an accurate diagnosis, select appropriate actions and support student learning. These inferences are, however, beyond the scope of the current chapter. For more information on how to validate inferences in an embedded formative assessment, see Hopster-den Otter et al. (2019).

2 Task Inference: How to Link the Domain of Interest to Observed Writing Performance

Inferences from the domain of interest to a person's writing performance are based on the assumption that the assessment tasks are relevant for and representative of the construct that one aims to measure, and hence, capture the desired performance. This inference forms the fundamental basis of writing assessment and requires a clear definition and operationalization of the construct of writing.

2.1 *Define the Construct of Writing That Is Central in the Intervention Study*

The need to theoretically define the construct of interest as a basis for task development and validation has received strong support in the field of language testing (Bachman, 1990; Chapelle, 2010). However, providing a clear definition of a complex construct like writing ability is not easy. It can be defined from many different perspectives on writing and writing development, such as (socio)cognitive, sociocultural, or (psycho)linguistic theories of writing (MacArthur et al., 2017, p. 2). It can also be defined in either a broad or narrow

sense (Rijlaarsdam et al., 2012). For instance, writing can be regarded as a general ability that is expected to lead to a relatively consistent performance across tasks and contexts, but it can also be limited to subskills of writing (e.g., planning or revision skills) or genre- or language-specific aspects of writing (e.g., argumentation structure or grammatical correctness).

How writing is defined in an intervention study determines the extent to which certain tasks can be considered as appropriate for measuring the construct. For example, in the 1950s the Educational Testing Service started to include indirect measures in the assessment of writing, such as a revision task or multiple-choice questions about a text (Huddleston, 1954). Even though these indirect measures lead to more reliable scores which might even be validly interpreted as indicators of subskills of writing, researchers showed that they did not fully capture students' overall writing performance (Godshalk et al., 1966). This implies that the construct of writing should not be defined too narrowly, as this limits the interpretations and decisions one can make. A too broad definition of writing ability also seems problematic, as research has consistently shown that writing performance highly fluctuates between tasks (Bouwer et al., 2015; Schoonen, 2012).

Therefore, researchers argued for an interactionalist view on writing, in which writing ability is regarded in interaction with, or depending on, the context of the performance (Bachman, 1990; Chapelle et al., 2010; Chalhoub-Deville, 2003). In this view, the definition of writing ability should be narrowed down to the communicative effectiveness of language use in a particular communicative context. This implies that the domain of interest within an intervention is bound by the communicative context in which the writing takes place, and that the possible universe of tasks that can be included in the assessment are thus characterized by the audience and communicative goals within the intervention. A related, but more pragmatic approach to defining the domain of interest is by specifying how assessment scores are interpreted and used within the intervention study (Kane, 2011). This links the definition of writing to the overall aim of the writing intervention. For instance, when the intervention aims to improve students' overall writing proficiency this calls for a rather broad definition of writing, but when it is aimed at improving specific subskills of writing, a narrow definition will suffice.

2.2 *Operationalize the Construct of Writing in Terms of Relevant Task Characteristics*

The construct definition can be used to operationalize writing ability in terms of task characteristics that will elicit the relevant writing performance. Following the interactionalist view on writing, this operationalization should not

only include a specification of what to write (about), but it should also explicitly specify the intended audience and communicative goal, since a language task is 'an activity that involves individuals in using language for the purpose of achieving a particular goal or objective in a particular situation' (Bachman & Palmer, 1996, p. 44). For example, instead of only asking students to write a short argumentative letter about bullying, the task should also clarify for whom students should write this text (e.g., for their peers) and with what communicative goal (e.g., to persuade). Ideally, the reader of the text is real (and not imagined) and the purpose for writing should be personally relevant to students and also exist in the world outside school or the assessment context. Such authentic and meaningful tasks will enhance students' motivation to write high-quality texts, which is important for valid conclusions about the intervention's effects on students' writing proficiency.

To ensure that the tasks are eliciting only construct-relevant performance, and that this performance is not contaminated by irrelevant constructs (Chapelle, 2011; Shaw et al., 2012), a careful analysis should be made of all other aspects that might affect performance on the writing tasks, such as the length of the text, knowledge of the topic, and the use of source material (for instance, in the case of integrated writing, see Knoch & Sitajalabhorn, 2013). To exclude or control for irrelevant aspects, specific attention should be paid to the conditions under which the performance will occur, such as writing with or without time constraints, the complexity of the sources, using a computer or paper-pencil, or the motivation for writing.

2.3 *Collect Validity Evidence to Substantiate How Performance on the Writing Tasks Reflects the Intended Construct of Writing*

Several methods can be used to evaluate whether the tasks elicit relevant and representative writing performance that reflects the domain of interest (Shaw et al., 2012). First, a pilot can be used to get insights into how students perform on the selected writing tasks (e.g., in terms of timing, cognitive constraints and/or demands). This method is only adequate if it is explicitly stated in advance what validation information is needed from such a (pilot) analysis and why this information justifies the use of the selected tasks in the intervention study as a measure of writing (ability). Second, experts could perform a content analysis of the writing tasks. This provides face or content validity evidence (Kane, 2011). Experts can be writing researchers or teachers who have ample expertise in writing education for the target group in terms of age and background (e.g., special needs, native or nonnative speakers). Based on their expertise, they can indicate the extent to which the selected tasks would elicit the desired writing performance. Finally, afterwards, statistical analysis

such as Rasch or factor analysis can be performed on performance data in order to explore the relationships between tasks, and to check for test bias and construct-irrelevant variance.

3 Scoring Inference: How to Link Writing Performance to Text Quality Scores

Students' performances on the selected writing tasks are quantified by providing scores to the quality of texts using a scoring procedure. It is generally assumed that the variances in scores are an appropriate and consistent indicator of differences in the quality of the observed writing performances. However, to which extent is this assumption warranted? One can make different choices in how text quality is scored, about the use of automatic measures or the selection of (human) raters, the kind of support or training that raters receive, or the scale that raters use for scoring text quality. Each of these choices has been, and still is, the subject of intense debate within the assessment community and the literature frequently presents contradictory findings regarding the reliability and validity of scores due to scoring method, rater background or expertise, and training of raters. Additionally, one should take into account that decisions on the scoring procedure, the trait-to-be-rated, the type of rater and the training one provides to raters are intertwined (cf. Barkaoui, 2007b, 2010; Bouwer et al., in press; Schoonen, 2005, 2012; Van den Bergh et al., 2012). To make sure that text quality scores can be validly interpreted as indicators of students' writing performance, researchers are advised to set up an assessment procedure with this complex interplay of factors in mind, weigh the pros and cons of every decision and document every step in the process.

3.1 *Decide If Text Quality Is Rated by Human Raters and/or by Automatic Analyses*

The decision for using automatic evaluation or human raters depends largely on the construct one aims to measure. Especially when the construct is narrowed down to particular linguistic aspects, so-called automated writing evaluation (AWE) may be a valid way to score writing performances. For example, AWE systems such as Coh-metrix (McNamara et al., 2014), SCA (Lu, 2010), TAASSSC (Kyle, 2016), TAALES (Kyle & Crossley, 2015) and Range software (Nation, 2021) offer various analytical measures for syntactic complexity and lexical density. These automatic measures appear to be largely consistent with human ratings, allowing for efficient and reliable scoring of linguistic features (Polio & Yoon, 2018).

All-encompassing AWE software such as e-rater (Attali & Burstein, 2006) also provide automatic indices of content-related features. However, for evaluating meaning and communicative effectiveness of texts, AWE systems are not comparable to human raters. Even though significant technological advances have been made, for instance the technology can now pretty accurately determine the degree to which a writer is "on-topic", the technology cannot yet adequately indicate whether a writer has constructed a "good" argument. This could lead to construct-underrepresentation in the scores (Shermis et al., 2017). Therefore, if automatic software is used, argue what measures are included and why they lead to construct-representative scores.

3.2 *Select Experienced Raters Based on Their Background and Expertise*

If human raters are involved, precautions have to be taken to avoid so-called rater effects (Myford & Wolfe, 2003). After all, we want differences in text scores to reflect variability in ratee performance instead of rater variability. Rater effects can have multiple causes. For instance, it can be due to different aspects that raters take into account when scoring text quality, or to halo effects in which a specific aspect of the text, or even of the ratee, radiates to the evaluation of other text quality features and/or the global evaluation of the writing product. Raters can also vary in how they use a scoring scale (Leckie & Baird, 2011). For instance, some raters are more severe than others which results in generally lower scores (i.e., severity effect), or their scores are restricted to the mid-range of a scoring scale (i.e., central tendency). This latter kind of rater effects are not problematic for most intervention studies, as text quality scores are generally used in a relative manner, that is, they are compared to each other in order to estimate the magnitude of improvements in writing of students. However, if scores are used as grades to make absolute decisions about the level of students' writing performance (cf. in norm-referenced and criterion-referenced testing; Bachman, 1990), these rater effects should be taken into account, for instance by using standardized scores instead of raw scores.

To control for rater effects, researchers can select raters based on their educational or professional background and their rating expertise. Research demonstrated that experienced raters are less susceptible to rater effects, and vary less in their approach to the scoring task (Wolfe et al., 1998) and the scores they assign accordingly (Barkaoui, 2010). As a result, intra- and inter-rater reliability is frequently higher among experts as opposed to novices (Leckie & Baird, 2011; Schoonen et al., 1997). Also, it is suggested that experienced raters and novices differ in the importance they attach to specific criteria (Barkaoui, 2010).

The selection of experienced raters is thus part of the validity argument. However, a question which emerges is how experience or expertise should be

defined in intervention studies as they are quite frequently used interchangeably. Lim (2011) proposes to reserve expertise to the quality of one's rating performance. Rater experience can be defined as having a teaching background and/or experience in rating texts for that particular age group. As it is plausible that such experienced raters are capable of translating the quality of performances to corresponding scores, it can be assumed that these raters' scores adequately reflect text quality.

The decision for experienced or novice raters can also depend on the task or trait to be scored. Studies have shown that content and structure can be reliably rated by both experienced or inexperienced raters, whereas language usage is most reliably rated by experts (Schoonen et al., 1997, 2005). Also, as tasks become more restricted (e.g., interlinear revision tasks), non-experts are as reliable as expert raters. Differences in rater experience or background may also become less of an issue with training (Weigle, 1994) and holistic rating scales (Schoonen, 2005). In either case, researchers should clearly define the type of rater that is included, based on their previous rating experience and educational or professional background, and they should argue why these raters' scores can be validly interpreted and used.

3.3 *Provide Raters with Clear Rating Instructions and Benchmark Examples*

To ensure that text quality scores align with the domain of interest, it is recommended to provide raters with clear instructions on how to rate text quality (cf. Barkaoui, 2010). There are several methods by which text quality can be rated, and researchers need to argue why the selected rating method is appropriate within the context of their intervention study. The two most known rating methods are holistic and analytic ratings. In holistic rating methods, raters provide a single score to the text-as-a-whole, either with or without predefined rating criteria (Charney, 1984). In most holistic rating methods, text quality scores reflect both content- and language-related aspects, but they could also reflect only a primary trait such as the extent to which the text successfully accomplished the rhetorical purpose of the task (Lloyd-Jones, 1977). In analytic rating methods, such as criteria-lists, rubrics, or checklists (see Jonsson & Svingby, 2007; Weigle, 2002), text features are measured separately. When analytic scores are combined into one score, it can also be used as an indicator of global text quality.

The choice for an analytic or a holistic rating method should be determined by the construct of writing one intends to measure. If one wants to have detailed information about specific features, an analytic rating method may be preferred. However, for global text quality evaluation both analytic and

holistic rating methods can be used, each with their pros and cons. The general assumption is that because analytic rating methods are more restrictive to readers they will result in higher inter-rater reliability (Wesdorp, 1981). A meta-analysis by Jonsson and Svingby (2007) on the use of analytic and holistic rubric use in complex performance assessment illustrates the acceptable reliability of (topic-specific) analytic rubrics with examples combined with rater training. However, think-aloud studies revealed that raters still vary in how they interpret and use analytic rubrics, especially with regard to the higher-level criteria (Barkaoui, 2007b; Lumley, 2002). Holistic rating, on the other hand, even though heavily criticized for its lack of reliability, has been associated with higher construct representation than analytic rating (Jonsson et al., 2021; Sadler, 2009). This raises the question whether the focus on rater reliability in analytic ratings does not go at the cost of its validity, especially against the background of a movement favoring the authenticity in variable reader responses to texts (Barkaoui, 2007a; Huot, 1990; Rijlaarsdam et al., 2012). The literature agrees on the fact that both analytic and holistic rating scales should ideally be validated in prior research that provides insights into how raters interpret and use the rating scales, and they should be accompanied not only by clear scoring guidelines and criteria but also by examples (Cumming et al., 2002).

Another method of evaluation which has shown its reliability and validity in quite a few state-of-the-art intervention studies in writing research is comparative evaluation. Instead of providing absolute scores to a text as in holistic or analytic rating methods, raters compare texts either with example texts from a corpus representing a specific text quality score or performance level (i.e., benchmark rating procedure, Bouwer et al., 2018; Bouwer & van den Bergh, in press; De Smedt & Van Keer, 2018; Limpo & Alvez, 2017; Raedts et al., 2017; Vandermeulen et al., 2020) or with other texts in the sample (i.e., comparative judgment, Lesterhuis et al., 2016; Pollitt, 2012; Verhavert et al., 2019). Comparative evaluation is a viable and valid alternative for absolute scoring as in analytic (Coertjens et al., 2017; Schoonen, 2005) or holistic scoring methods (Bouwer & van den Bergh in press). Not only is it easier for raters to compare two texts than to assign a single score to a text but it also prevents norm-shifting by raters (Lesterhuis et al., 2016). The two comparative methods can also be integrated, in which benchmarks are first carefully selected based on the results of a comparative assessment, after which the benchmarks can be used to assess a second and larger set of texts (e.g., Bouwer et al., in press; De Smedt et al., 2020; McGrane et al., 2018; Vandermeulen et al., 2020). This two-stage process increases the quality of the selected benchmarks, and hence, promotes the validity of benchmark ratings (Osborn Popp et al., 2009). It has also been demonstrated that benchmark scales, which

are usually task-dependent, can be used for rating similar tasks on a different topic (Bouwer & van den Bergh, in press; Tillemal et al., 2013).

3.4 *Organize Training for Raters to Rate, Compare and Discuss Example Texts*

Rater training usually involves bringing raters together to collectively rate, compare and discuss various kinds of example texts, both clear-cut and difficult-to-score texts, using the predefined scoring criteria, categories or scale points (Roch et al., 2012; Wang et al., 2017; Wolfe et al., 2016). From the literature on rater training emerges mainly that such a training provides a frame of reference to raters, may reduce rater biases and result in a shared understanding and more consistency between raters (Bachman & Palmer, 2010; Lumley, 2002; Shohamy et al., 1992; Wang et al., 2017; Weigle, 1994). Recently, it has also been demonstrated that by collectively comparing a series of example texts of varying performance levels raters can learn how to conceptualize text quality (Van Gasse et al., 2019).

Research, however, also illustrates that training does not guarantee that raters interpret and apply scoring criteria in the same manner. Trained raters may still differ in focus, weighting and decision-making processes (Eckes, 2008; Vaughan, 1991). These different rater perspectives should rather be embraced than avoided, as they are essential for understanding the complexity and multidimensionality of text quality and sharing norms and standards (Jølle, 2015; Van Gasse et al., 2019). Key is then to make room for comparing and discussing different norms and perspectives on the evaluation of text quality. Researchers should therefore make a thoughtful decision about the training specifics (i.e., format, modality, duration, examples) to be able to justify score validity without "pressur[ing] readers into agreement" Hamp-Lyons, 2007, p. 3) all the while striving for "the formation of an assenting community that feels a sense of ownership of the standards and the process" (White, 1985, p. 69, as cited by Hamp-Lyons, 2007, p. 5).

3.5 *Collect Validity Evidence to Substantiate Whether Text Quality Is Scored as Intended*

A final part of the validity argument for the scoring inference is to justify that raters score text quality as intended. After all, providing raters with scoring instructions, example texts and sufficient training does not guarantee that they follow the scoring procedure adequately and consistently over time. Therefore, researchers need to monitor and evaluate whether raters followed the intended rating procedure and how they applied scoring criteria, guidelines, and/or benchmark examples when scoring text quality. This could be done by

collecting additional validity evidence, such as think-aloud data of raters, or by estimating inter- and intra-rater reliability.

4 Generalizability Inference: How to Generalize from Text Quality Scores to Writing Ability

In writing interventions, researchers generally interpret text quality scores on a sample of tasks as the students' expected writing performance generalized over tasks and raters. But are these conclusions always warranted? For instance, can the average scores for three argumentative essays holistically rated by two raters be used as an indicator of students' overall proficiency in argumentative writing? When is it appropriate and plausible to consider students with high text quality scores as good (argumentative) writers, and students with low scores as writers who are less proficient? The generalization from task performance to performance in the larger domain is only justified if the scores are representative of this performance domain and if the sample is large enough to control for sampling error (Kane, 2006). To validate such inferences, researchers have to argue that the writing tasks adequately sample the intended construct of writing ability, and they have to provide evidence that text quality scores can be generalized to students' writing performance across tasks and raters.

4.1 *Have More Than One Rater Who Rates Text Quality*

As discussed in Section 3, the potentially considerable variability between raters in how they evaluate text quality may form a serious threat to valid interpretations of text quality scores. Conditions such as having only experienced raters who receive training and strict scoring protocols will only improve the reliability of scores to a limited extent. At least for the higher-order aspects of text quality, there will always be some variance between raters, due to the subjectivity in the rating process and individual perspectives on what one considers as a good text (cf. schools-of-thought, Diederich et al., 1987; Lesterhuis, 2018). This makes generalizations of scores based on only one rater hardly plausible. It is, therefore, commonly advised to have multiple raters or rater panels rate text quality (Gebril, 2009; Schoonen, 2005; Weigle, 2002).

In some studies, especially when many texts need to be rated such as in a large-scale national sample as in the LIFT-project by Vandermeulen et al. (2020) or in the large-scale intervention study of Bouwer et al. (2018), a design of overlapping rater teams is used to generalize across multiple raters. In this design, the writing products are randomly divided into subsamples that equal

the number of raters. Each rater receives three subsamples according to a design in which there is systematic overlap between raters. This overlapping design allows for the estimation of individual and jury rater reliabilities (Van den Bergh & Eiting, 1989). In general, jury reliabilities for scores that are averaged across two or three raters are shown to be higher than individual rater reliabilities. Jury scores are also considered to be more valid, as they include different perspectives on text quality, which justify generalizations over raters (Lesterhuis, 2018).

4.2 *Have More Than One Writing Task at Each Measurement Occasion*

Tasks are another source of variability in the writing assessment, indicating that students do not perform consistently across tasks (Bouwer et al., 2015; Godshalk et al., 1966; Schoonen, 2005, 2012; Van den Bergh et al., 2012). Individual differences in text quality scores across tasks can be explained by effects of both topic and genre knowledge (Bouwer et al., 2015). This means that a single text does not provide a reliable estimate of overall writing proficiency. In fact, it is nothing more than a one-item test, and even though it is a large item that is full of information, the performance on a single task cannot be generalized to performance on other tasks. Thus, to allow for inferences on students' overall writing proficiency across tasks, one needs to collect multiple texts written at each measurement occasion based on tasks that vary in topic and genre.

4.3 *Provide an Estimation of the Generalizability of Text Quality Scores in Your Study*

Ultimately, this brings us to the question: how many writing tasks and how many raters should be part of the assessment to warrant generalizations to students' writing proficiency? Previous generalizability studies have shown that tasks and task-related interactions generally explain more of the variability in writing scores than raters and rater-related interactions (Gao & Brennan, 2001; Kim et al., 2017;). Also, in some studies the variance due to the student-by-task interaction is larger than the variance due to students (Bouwer et al., 2015; Lehman, 1990). This means that in order to control for the variance due to raters and tasks, and hence to allow for generalizations, relatively more tasks than raters are needed.

The decision for the exact number of tasks and raters depends on the context and purpose of the assessment. When the aim is to generalize to genre-specific writing, it is advised to include at least three to five tasks with two to three raters to the assessment (Kim et al., 2017; Schoonen, 2005; 2012; Van den Bergh et al., 2012). For generalizations beyond genre, students should at least write three texts in each of four different genres, such as argumentative, narrative,

descriptive and personalized texts (Bouwer et al., 2015). The generalizability is also affected by grade. For instance, Van den Bergh et al. (2012) showed that university students perform more consistently in writing than 9th-grade students, which means that fewer tasks are needed to allow for generalizations about the writing performance of university students compared to college students. With regard to the effects of the rating procedure, multiple studies have shown that the generalizability of scores is higher when texts are rated holistically instead of analytically. This can be explained by the task-dependency of analytic criteria or rubrics, which is associated with higher task-related variance (Schoonen, 2005; Van den Bergh et al., 2012). A study by Bouwer and van den Bergh (in press) showed that the generalizability was even higher for benchmarks ratings, as they were associated with less rater variance than holistic ratings as well as with less task-specific variance than analytic ratings. In addition, the generalizability is higher for language-related features than for features related to content and organization (Schoonen, 2005).

As generalizability coefficients in writing assessments are likely to vary from one sample to another (Gao & Brennan, 2001, p. 192), we advise researchers to provide an estimation of the generalizability of text quality scores in their own study. For instance, by using the intraclass correlation as an indicator for the generalizability of scores across raters, or by estimating the magnitude of the variance components due to students, tasks, genres, and raters using multilevel modeling (for more information, see chapter 14 by Van den Bergh and De Maeyer, 2023).

5 Extrapolation Inference: How to Extrapolate from the Assessment to Writing in General

The previous steps focused on how to make evidence-informed decisions in the assessment design and to support these decisions with clear arguments. The fourth step of extrapolation does not necessarily refer to the assessment design, but rather to the broader domain to which we want to make inferences. This is closely related to the first step of task selection, in which the boundaries of the construct of interest were specified and operationalized by specific characteristics of writing tasks and performances. In writing interventions, however, scores on tasks within an assessment context are often used to make broader claims about writing within educational practice. If this is the aim, it is important to critically examine and argue whether scores can be extrapolated to the broader practice domain with respect to effects over time, the defined skill and the measurement situation.

5.1 *Include Delayed Posttests or Longitudinal Measures to Justify Maintenance Effects*

As researchers, we often assume that the effects of interventions are maintained over time, that is, when students show a certain level of writing performance at the end of an intervention, we expect them to also show this level of performance at a later occasion. However, Rijlaarsdam et al. (2012) state that students' writing performance on one occasion does not always predict how they will perform on similar tasks in the future. Roger & Graham's (2008) meta-analysis also shows a difference between scores on posttests and on delayed posttests. Thus, in order to make claims about improvements in students' writing performance beyond the immediate assessment, and to extrapolate the findings over time, it is essential to follow a longitudinal measurement design or to add a delayed posttest to the intervention. Rogers & Graham (2008) suggest that every measurement that takes place after three weeks after the intervention provides important information on maintenance effects. Another way to investigate maintenance effects of the intervention is by using a switching replication design (Bouwer et al., 2018), see also chapter 14 in the current volume by Van den Bergh and De Maeyer, 2023.

5.2 *Measure Performance on Other Genres or Subskills to Justify Transfer Effects*

Researchers sometimes draw conclusions that go beyond the specified domain of interest. That is, they claim that students' acquired writing skills will extrapolate to domains or tasks that are not part of the assessment. This extrapolation inference should not be confused with generalization of scores to tasks and raters that are similar to the ones used in the assessment. For example, when the aim of an intervention is to enhance planning skills within the context of argumentative writing, can we assume that these planning skills transfer to other genres as well? Again, empirical evidence is needed to warrant this assumption. Dostal and Wolbers (2016), for example, show how strategic and interactive writing instruction within a narrative genre also improves students' scores on writing information reports. Gentil (2011) argues that genre knowledge in one language can be transferred to another language.

In intervention research, the reasoning can also be reversed. De Smedt et al. (2020), for example, developed an intervention specifically aimed at improving students' performance on descriptive writing tasks. In the intervention, students were taught general planning, composing, and revising strategies, however, they were not explicitly instructed to use these strategies also for writing in the other genres. They measured both descriptive and narrative writing before and after the intervention, in order to investigate not only the

effects on descriptive writing, but also whether spontaneous transfer to narratives occurred, and found that this was not the case. Based on these findings, they attributed students' improvements in descriptive writing to the explicit strategy-based instructions and stated that transfer did not take place because students were not stimulated to do so.

5.3 *Evaluate the Representativeness and Authenticity of the Assessment Tasks to Justify Conclusions to Real-Life Writing*

Researchers may also want to extrapolate their conclusions regarding the effectiveness of the intervention to writing in the real world. In this case, the aim of the intervention is not only to improve students' writing performance in a specific genre within the context of the classroom (e.g., writing a motivation letter for a teacher), but also to affect students' writing performance in other courses or in their life outside of school (e.g., writing an application letter). However, as it is hardly possible to investigate the extent to which newly-acquired writing skills also transfer to writing in the real world (after all, this would always involve assessment-based tasks, cf. Bachman, 1990), it is not yet known to what extent this extrapolation inference is justified. A possible way to investigate whether it is plausible that the assessment scores of an intervention can be extrapolated to a broader practice domain, is to ask teachers or other relevant experts to evaluate the extent to which (the) tasks are relevant and representative of future performance in the curriculum and beyond (Shaw et al., 2012).

6 Decision Inference: How to Make Decisions on the Effectiveness of an Intervention

A final aim of researchers is to draw conclusions about the effects of the intervention on students' writing ability. In this section, we will briefly discuss the assumptions behind the decision-making process: what do we mean when we claim that an intervention is effective and when are such decisions justified?

6.1 *Evaluate the Intervention Design and Decide What Conclusions Are Warranted*

The conclusions that can be made regarding the effectiveness of the intervention depend on the research design. For instance, a classic pretest-posttest design allows researchers only to draw conclusions about improvements in students' writing performances after the intervention. To also attribute improvements in students' writing to the intervention, and not merely to progression over time, their performance should be compared to that of students in a con-

trol condition. This control condition could be a group of students who follow their regular writing practices (i.e., business-as-usual condition, e.g., De Smedt et al., 2020) or who receive another form of instruction (e.g., López et al., 2021). For more information on (quasi-)experimental designs in writing intervention research, see Graham and Harris (2014) or chapter 14 in the current volume by Van den Bergh and De Maeyer, 2023.

6.2 *Provide Meaningful Effect Sizes That Indicate the Degree of Mastery in Writing*

Key is not only to report whether there is an effect, but also what the strength of the effect is. As prescribed by AERA et al. (2014), these effect size measures should be paired with indices reflecting the degree of uncertainty (e.g., standard errors, confidence intervals), if they are used to draw inferences that go beyond describing the sample from which the data have been collected. Regular effect sizes, however, only provide information regarding the relative writing performance of students at different measurement occasions or in different conditions (i.e., norm-referenced testing). They do not reveal anything about the level of writing that students can master after the intervention (i.e., criterion-referenced testing). To allow for more useful interpretations of assessment outcomes, Lipsey et al. (2012) and Kraft (2020) propose to report effect sizes as a function of progression by number of months, changes in percentile rank, achievement gaps, differences between teachers or schools, or by referring to national reference frameworks or the Common European Framework of References. An example of such a more intuitive interpretation is presented by Bouwer et al. (2018), who compared the improvement of upper-elementary students' writing after the intervention to the general improvement in writing between grade 4 to 6. An even more innovative way of expressing effects in terms of grade progression is by comparing the intervention effects to a baseline sample using Bayesian statistics. Van den Bergh and colleagues (2023) were able to estimate the difficulty of each writing task for students in different grades using a Bayesian approach. They used these priors to relate the effect of the intervention to regular progression across grade, while taking into account the difficulty of the tasks.

6.3 *Collect Process Data to Reveal How the Intervention Affects Writing*

There are many factors that can affect the effectiveness of the intervention. For instance, interventions may only work well for some students, in some contexts, or only with particular tasks. Therefore, Rijlaarsdam et al. (2017) argue that researchers should provide an insight into what is working for whom and how. The 'how' is related to the theoretical foundations of the intervention and

why an intervention is expected to change students' underlying writing processes. The relationship between the writing process and the outcomes of that process in terms of observed performances on the writing tasks should not only be explicated by the design principles of the intervention (see chapter 3 by De Smedt, 2023), but they should ideally also be measured as part of the writing assessment (see chapter 11 by Vandermeulen, 2023). This information is crucial for making valid decisions about the effectiveness of the intervention as well as for further theory-building.

7 Discussion

Assessing writing in a valid and reliable manner is a complex endeavor, especially for researchers whose primary focus is on the teaching of writing. To support writing intervention researchers, we have applied an existing validation framework by Kane to writing assessment in intervention studies. Following this framework, we have argued that the validity of the use and interpretation of writing assessment scores depends on the plausibility of five inferences, or steps, in the claim from the observed writing performance to decisions on the effectiveness of the intervention: selecting writing tasks, scoring text quality, generalization, extrapolation and decision-making. In this chain of inferences, each step needs to be justified by theoretical or empirical evidence, because if any inference fails, the whole validity argument fails. By following the proposed guidelines in this chapter, researchers can systematically substantiate their conclusions regarding the effectiveness of the intervention and counterargue possible threats to validity already at an early stage of the intervention.

Even though all five validation steps seem to be important in every intervention study, the emphasis given to each step in the argumentation may depend on the study's context and purpose. This means that researchers have to decide what the most critical or challenging inferences or assumptions are when interpreting or using writing assessment information. In line with Kane et al. (1999), we argue that the most attention in writing intervention studies should be paid to the generalizability of the assessment outcomes over raters and tasks, and to the extrapolation of the scores over time and to other writing domains. After all, is the aim of intervention research not to understand how students' overall writing skills can be improved, not only for a particular task or within a specific assessment context? In most intervention studies, however, these inferences seem to be challenged, as writing is often measured with only one task and one rater, and intervention designs rarely include delayed posttest measures to test maintenance effects or a different type of task to test transfer effects (cf.

Graham & Harris, 2014). As a consequence, decisions regarding the effectiveness of the interventions cannot be generalized over tasks, raters, occasions or contexts. Therefore, we advise researchers to at least argue whether the generalization and extrapolation beyond the assessment domain is warranted. We also suggest that decisions in the other steps of the validation framework, such as the selection of tasks and rating procedures, should be made with regard to their effects on the generalizability of the results. The framework that we present in this chapter will not solve the complexity of writing assessment designs, as there is not a single way of guaranteeing a qualitative writing assessment for every study, but we hope that we have provided researchers with the necessary knowledge and useful guidelines to make well-structured and evidence-informed decisions for assessing writing in their intervention studies.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Arrimada, M. (2023). Response to intervention as a framework to implement writing interventions: opportunities and challenges. In F. De Smedt, R. Bouwer, T. Limpo & S. Graham (Eds.), *Conceptualizing, Designing, Implementing and Evaluating Writing Interventions.* (*Studies in Writing Series*) (pp. 175–198). The Netherlands: Brill Publishers.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *The Journal of Technology, Learning, and Assessment, 4*(3), 1–30.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L.F. & Palmer, A.S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Barkaoui, K. (2007a). Participants, Texts, and Processes in ESL/EFL Essay Tests: A Narrative Review of the Literature. *The Canadian Modern Language Review / La revue canadienne des langues vivantes*, *64*(1), 99–134

Barkaoui, K. (2007b). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86–107. https://doi.org/10.1016/j.asw.2007.07.001

Barkaoui, K. (2010). Variability in ESL Essay Rating Processes: The Role of the Rating Scale and Rater Experience. *Language Assessment Quarterly*, 7(1), 54–74. https://doi.org/10.1080/15434300903464418

Bouwer, R., Béguin, A., Sanders, T., & Van den Bergh, H. (2015). Effect of genre on the

generalizability of writing scores. *Language Testing, 32*(1), 83–100. https://doi.org/10.1177/0265532214542994

Bouwer, R., Koster, M., & Van den Bergh, H. (2018). Effects of a strategy-focused instructional program on the writing quality of upper elementary students in the Netherlands. *Journal of Educational Psychology, 110*(1), 58–71. https://doi.org/10.1037/edu0000206

Bouwer, R., Koster, M., & van den Bergh, H. (in press). Benchmark rating procedure, best of both worlds? Comparing procedures to rate text quality in a reliable and valid manner. *Assessment in Education: Principles, Policy & Practice*.

Bouwer, R., Lesterhuis, M., De Smedt, F., Van Keer, H., & De Maeyer, S. (in press). Comparative approaches to the assessment of writing: Reliability and validity of benchmark rating and comparative judgement. *Journal of Writing Research*.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*(4), 369–383.

Chapelle, C.A., Enright, M.K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practices, 29*, 3–13. https://doi.org/10.1111/j.1745-3992.2009.00165.x

Chapelle, C.A. (2011). Validity argument for language assessment: The framework is simple … *Language Testing, 29*(1), 19–27. https://doi.org/10.1177/0265532211417211

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English*, *18*(1), 65–81.

Coertjens, L., Lesterhuis, M., Verhavert, S., Van Gasse, R., & De Maeyer, S. (2017). Teksten beoordelen met criterialijsten of via paarsgewijze vergelijking: een afweging van betrouwbaarheid en tijdsinvestering. [Evaluating texts with criteria or comparative evaluation: a comparison of reliability and time]. *Pedagogische Studiën*, *94*(4), 283–303.

Cumming, A., Kantor, R., & Powers, D.E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal*, *86*(1), 67–96. https://doi.org/10.1111/1540-4781.00137

De Smedt, F. (2023). Designing and reporting interventions: From a blueprint to a systematic and analytic description. In F. De Smedt, R. Bouwer, T. Limpo & S. Graham (Eds.), *Conceptualizing, Designing, Implementing and Evaluating Writing Interventions.* (*Studies in Writing Series*) (pp. 37–53). The Netherlands: Brill Publishers.

De Smedt, F., Graham, S. & Van Keer, H. (2020). "It takes two": The added value of structured peer-assisted writing in explicit writing instruction. *Contemporary Educational Psychology*, *60*, https://doi.org/10.1016/j.cedpsych.2019.101835

De Smedt, F., & Van Keer, H. (2018). Fostering writing in upper primary grades: a study into the distinct and combines impact of explicit instruction and peer assistance. *Reading and Writing, 31*(2), 325–354. https://doi.org/10.1007/s11145-017-9787-4

Diederich, P.B., French, J.W., & Carlton, S.T. (1961). *Factors in judgments of writing ability* (Research Bulletin RB-61–15). Educational Testing Service.

Dostal, H.M. & Wolbers, K.A. (2016). Examining Student Writing Proficiencies Across Genres: Results of an Intervention Study. *Deafness & Education International, 18*(3), 159–169. https://doi.org/10.1080/14643154.2016.1230415

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155–185. https://doi.org/10.1177/0265532207086780

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly, 2*(3), 175–196. https://doi.org/10.1207/s15434311laq0203_1

Gao, X., & Brennan, R.L. (2001). Variability of Estimated Variance Components and Related Statistics in a Performance Assessment. *Applied Measurement in Education*, *14*(2), 191–203. https://doi.org/10.1207/s15324818ame1402_5

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing, 26*(4), 507–531.

Gentil, G. (2011). A Biliteracy Agenda for Genre Research. *Journal of Second Language Writing, 20*(1), 6–23.

Godshalk, F.I., Swineford, F., & Coffman, W.E. (1966). *The measurement of writing ability*. College Entrance Examination Board.

Graham, S. & Harris, K.R. (2014). Conducting high quality writing intervention research: Twelve recommendations. *Journal of Writing Research 6* (2), 89–123. https://doi.org/10.17239/jowr-2014.06.02.1

Hamp-Lyons, (2007). Editorial. *Assessing Writing*, *12*(1), 1–9. https://doi.org/10.1016/j.asw.2007.05.002

Hopster-den Otter, F., Wools, S., Eggen, T.J.H.M., & Veldkamp, B.P. (2019). A general framework for the validation of embedded formative assessment. *Journal of Educational Measurement, 56*(4), 715–732. https://doi.org/10.1111/jedm.12234

Huddleston, E. (1954). Measurement of writing ability at the college-entrance level: Objective vs. subjective testing techniques. *The Journal of Experimental Education, 22*, 165–213.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know, what we need to know. *College Composition and Communication*, *41*(2), 201–213.

Jølle, L. (2015) Rater strategies for reaching agreement on pupil text quality. *Assessment in Education: Principles, Policy & Practice*, *22*(4), 458–474. https://doi.org/10.1080/0969594X.2015.1034087

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Jonsson, A., Blan, A., & Hartell, E. (2021). Analytic or holistic? A study about how to

increase the agreement in teachers' grading. *Assessment in Education: Principles, Policy, & Practice*. https://doi.org/10.1080/0969594x.2021.1884041

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.

Kane, M. (2011). Validating score interpretations and uses: Messick lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing, 29*(1), 3–17.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5–17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x

Kim, Y.-S.G., Schatschneider, C., Wanzek, J., Gatlin, B., & Otaiba, S.A. (2017). Writing evaluation: rater and task effects on the reliability of writing scores for children in Grades 3 and 4. *Reading and Writing*, *30*(6), 1287–1310. https://doi.org/10.1007/s11145-017-9724-6

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writers online: How does it compare with face-to-face training? *Assessing Writing*, *12*, 26–43. https://doi.org/10.1016/j.asw.2007.04.001

Knoch, U., & Sitajalabhorn, W. (2013). A closer look at integrated writing tasks: Towards a more focussed definition for assessment purposes. *Assessing Writing*, 18, 300–308.

Kraft, M.A. (2020). Interpreting effect sizes of education interventions. *Educational researchers, 49*(4), 241–253, https://doi.org/10.3102/0013189X20912798

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Unpublished doctoral dissertation]. Georgia State University

Kyle, K., & Crossley, S.A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757–786. https://doi.org/10.1002/tesq.194

Leckie, G., & Baird, J. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399–418.

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (2016). Comparative Judgment as a promising alternative to score competences. In E. Cano & G. Ion (Eds.), *Innovative Practices for Higher Education Assessment and Measurement* (pp. 119–138). IGI Global. https://doi.org/10.4018/978-1-5225-0531-0

Lesterhuis, M. (2018). *The validity of comparative judgement for assessing text quality. An assessor's perspective* [Unpublished doctoral dissertation]. University of Antwerp, Antwerp.

Lim, G.S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543–560. https://doi.org/10.1177/0265532211406422

Limpo, T., & Alves, R.A. (2017). Relating beliefs in writing skill malleability to writing

performance: The mediating role of achievement goals and self-efficacy. *Journal of Writing Research*, *9*(2), 97–125. https://doi.org/10.17239/jowr-2017.09.02.01

Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., Roberts, M., Anthony, K.S., Busick, M.D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms*. National Center for Special Education Research.

Lloyd-Jones, R. (1977). Primary-trait scoring. In C.R. Cooper, & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 33–68). National Council of Teachers of English.

López, P., Torrance, M., Rijlaarsdam, G., Fidalgo, R. (2021). Evaluating effects of different forms of revision instruction in upper-primary students. *Reading & Writing*, 34(7), 1741–1767. https://doi.org/10.1007/s11145-021-10156-3

Lu, X. (2010). Automatic measurement of syntactic complexity in child language acquisition. *International Journal of Corpus Linguistics*, *14*(1), 3–28.

Lumley, T. (2002) Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*(3), 246–274. https://doi.org/10.1191/0265532202lt2300a

McNamara, D.S., Graesser, A.C., McCarthy, P.M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

MacArthur, C.A., Graham, S. & Fitzgerald, J. (2017). *Handbook of writing research* (2nd ed.). The Guilford Press.

McGrane, J.A., Humphry, S.M., & Heldsinger, S. (2018). Applying a thurstonian, two-stage method in the standardized assessment of writing. *Applied Measurement in Education, 31*(4), 297–311. https://doi.org/10.1080/08957347.2018.1495216

Myford, C.M., & Wolfe, E.W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Nation, P. (2021). Vocabulary analysis program: Range [Computer software]. Victoria University of Wellington. https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-analysis-programs

Osborn Popp, S.E., Ryan, J.M., & Thompson, M.S. (2009). The critical role of anchor paper selection in writing assessment. *Applied Measurement in Education*, *22*(3), 255–271. https://doi.org/10.1080/08957340902984026

Polio, C., & Yoon, H. (2018). The reliability and validity in automated tools for examining variation in syntactic complexity across genres. *International Journal of Applied Linguistics*, *28*(1), 165–188. https://doi.org/10.1111/ijal.12200

Pollitt, A. (2012). The method of Adaptive Comparative Judgement. *Assessment in Education: Principles, Policy & Practice*, *19*(3), 281–300. http://doi.org/10.1080/0969594X.2012.665354

Raedts, M., Van Steendam, E., De Grez, L., Hendrickx, J., & Masui, C. (2017). The effects of different types of video modelling on undergraduate students' motivation and

learning in an academic writing course. *Journal of Writing Research, 8*(3), 399–435. https://doi.org/10.17239/jowr-2017.08.03.01

Rijlaarsdam, G., Van den Bergh, H., Couzijn, M., Janssen, T., Braaksma, M., Tillema, M., Van Steendam, E., & Raedts, M. (2012). *Writing*. In K.R. Harris, S. Graham, T. Urdan, A.G. Bus, S. Major, & H.L. Swanson (Eds.), *APA handbooks in psychology®. APA educational psychology handbook* (*Vol. 3. Application to learning and teaching*, pp. 189–227). American Psychological Association. https://doi.org/10.1037/13275-009

Rijlaarsdam, G., Janssen, T., Rietdijk, S., & van Weijen, D. (2017). Reporting design principles for effective instruction of writing: Interventions as constructs. In *Design principles for teaching effective writing* (pp. 280–313). Brill.

Roch, S.G., Woehr, D.J., Vipanchi, M., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology*, *85*(2), 370–395. https://doi.org/10.1111/j.2044-8325.2011.02045.x

Rogers, L.A. & Graham, S. (2008). A meta-analysis of single subject design writing intervention research. *Journal of Educational Psychology, 100* (4), 879–906. https://doi.org/10.1037/0022-0663.100.4.879

Sadler, D.R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*(2), 159–179. https://doi.org/10.1080/02602930801956059

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modelling. *Language Testing*, *22*(1), 1–30. https://doi.org/10.1191/0265532205lt2950a

Schoonen, R. (2012). The validity and generalizability of writing scores: The effect of rater, task and language. In E. Van Steendam, M. Tillema, G. Rijlaarsdam, & H. Van den Bergh (Eds.), *Measuring Writing: Recent Insights into Theory, Methodology and Practices* (pp. 1–22). Brill. https://doi.org/10.1108/S1572-6304(2012)0000027005

Schunk, D.H. & Swartz, C.W. (1993). Goals and Progress Feedback: Effects on Self-Efficacy and Writing Achievement. *Contemporary Educational Psychology*, *18*(3), 337–354. https://doi.org/10.1006/ceps.1993.1024.

Shaw, S., Crisp, V., & Johnson, N. (2012). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assessment in Education: Principles, Policy & Practice*, *19*(2), 159–176. https://doi.org/10.1080/0969594x.2011.563356

Shermis, M.D., Burstein, J., Elliot, N., Miel, S., & Foltz, P.W. (2017). Automated writing evaluation: An expanding body of knowledge. In C.A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 395–410). The Guilford Press.

Shohamy, E., Gordon, C.M. and Kraemer, R. (1992): The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal*, *76*, 27–33.

Tillema, M., van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2013). Quantifying the quality difference between L1 and L2 essays: A rating procedure with bilingual raters and L1 and L2 benchmark essays. *Language Testing*, *30*(1), 71–97. https://doi.org/10.1177/0265532212442647

Van den Berg, D., Vandermeulen, N., Lesterhuis, M., De Maeyer, S., Van Steendam, E., Rijlaarsdam, G. & Van den Bergh, H. (2023). How Prior Information from National Assessments can be used when Designing Experimental Studies without a Control Group. *Journal of Writing Research*, *14*(3), 447–469. https://doi.org/10.17239/jowr-2023.14.03.05

Van den Bergh, H. & De Maeyer, S. (2023). On multilevel modeling in writing research: an example. In F. De Smedt, R. Bouwer, T. Limpo & S. Graham (Eds.), *Conceptualizing, Designing, Implementing and Evaluating Writing Interventions.* (*Studies in Writing Series*) (pp. 293–313). The Netherlands: Brill Publishers.

Van den Bergh, H., De Maeyer, S., Van Weijen, D., & Tillema, M. (2012). Generalizability of text quality scores. In E. Van Steendam, M. Tillema, G. Rijlaarsdam, & H. Van den Bergh (Eds.), *Measuring Writing: Recent Insights into Theory, Methodology and Practices* (pp. 23–32). Brill. https://doi.org/10.1108/S1572-6304(2012)0000027005

Van den Bergh, H., & Eiting, M.H. (1989). A method of estimating rater reliability. *Journal of Educational Measurement*, *26*(1), 29–40.

Vandermeulen, N. (2023). Process measures as input for and as outcome of writing intervention studies. In F. De Smedt, R. Bouwer, T. Limpo & S. Graham (Eds.), *Conceptualizing, Designing, Implementing and Evaluating Writing Interventions.* (*Studies in Writing Series*) (pp. 226–252). The Netherlands: Brill Publishers.

Vandermeulen, N., De Maeyer, S., Van Steendam, E., Lesterhuis, M., Van den Bergh, H., & Rijlaarsdam, G. (2020). Mapping synthesis writing in various levels of Dutch upper-secondary education. A national baseline study on text quality, writing process and students' perspectives on writing. *Pedagogische Studien*, *97*(3), 187–236.

Van Gasse, R., Lesterhuis, M., Verhavert, S., Bouwer, R., Vanhoof, J., Van Petegem, P., & De Maeyer, S. (2019). Encouraging professional learning communities to increase the shared consensus in writing assessments: The added value of comparative judgement. *Journal of Professional Capital and Community*, *4*(4), pp. 269–285. https://doi.org/10.1108/JPCC-08-2018-0021

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111–125). Ablex.

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education*, *26*(5), 541–562. http://doi.org/10.1080/0969594X.2019.1602027

Wang, J., Engelhard, G., Raczynski, K., Song, T., & Wolfe, E.W. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing*, *33*, 36–47. https://doi.org/10.1016/j.asw.2017.03.003

Weigle, C.S. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197–223. https://doi.org/10.1177/026553229401100206

Weigle, S.C. (2002). *Assessing writing*. Cambridge University Press.

Wesdorp, H. (1981). *Evaluatietechnieken voor het moedertaalonderwijs* [Assessment methods for first language education]. Den Haag: Stichting voor Onderzoek voor het onderwijs.

Wolfe, E.W., Kao, C.W., & Ranney, M. (1998). Cognitive differences in proficient and nonproficient essay scorers. *Written Communication*, *15*(4), 465–492. https://doi.org/10.1177/0741088398015004002

Wolfe, E.W., Song, T., & Jiao, H. (2016). Features of difficult-to-score essays. *Assessing Writing*, *27*, 1–10. https://doi.org/10.1016/j.asw.2015.06.002