Annual Review of Applied Linguistics (2014), **34**, 123–153.
© Cambridge University Press, 2014, 0267-1905/14 $16.00
doi: 10.1017/S0267190514000063

Mixed Methods Research in Language Testing and Assessment

Eunice Eunhee Jang, Maryam Wagner, and Gina Park

As an alternative paradigm, mixed methods research (MMR), in general, endorses pluralism to understand the complex nature of a social world from multiple perspectives and multiple methodological lenses, each of which offers partial, yet valuable, insights. This methodological mixing is not limited to mixing of methods, but extends to the entire inquiry process. Researchers in language testing and assessment (LTA) are increasingly turning to MMR in order to understand the complexities of language acquisition and interaction among various language users, and also to expand opportunities to investigate validity claims beyond the three traditional facets of construct, content, and criterion validity. We use current conceptualizations of validity as a guiding framework to review 32 empirical MMR studies that have been published in LTA since 2007. Our systematic review encompassed multiple areas of foci, including the rationale for the use of MMR, evidence of collaboration, and synergetic effects. The analyses revealed several key trends including: (a) triangulation and complementarity were the prevalent uses of MMR in LTA; (b) the majority of the studies took place predominantly in higher education learning contexts with adult immigrant or university populations; (c) aspects of writing assessment were most frequently the focus of the studies (compared to other language modalities); (d) many of the studies explicitly addressed facets of validity, and others had significant implications for expanding notions of validity in LTA; (e) the majority of the studies avoided mixing at the data analysis stage by distinguishing data types and reporting results separately; and (f) integration occurred primarily at the discussion stage. We contend that LTA should embrace MMR through creative designs and integrative analytic strategies to seek new insights into the complexities and contexts of language testing and assessment.

MIXED METHODS RESEARCH IN LANGUAGE TESTING AND ASSESSMENT

The field of language testing and assessment (LTA) has been traditionally dominated by a quantitative paradigm[1]; however, it has experienced a significant shift in the past two decades in the wake of expansion of both the theoretical definitions of language competence and test validity. LTA practice has faced

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

prolific demand to serve various purposes from informing individual students' learning to leveraging educational policy and curricular reforms and accountability. As a consequence, researchers in LTA are increasingly turning to mixed methods research (MMR) to address the complexity of phenomena underlying LTA practices.

Simply speaking, MMR is an inquiry approach that includes both qualitative and quantitative methods within a single study or program of inquiry (Johnson, Onwuegbuzie, & Turner, 2007). The legitimacy of MMR as an inquiry approach in applied linguistics has been acknowledged (Dörnyei, 2007), with an increasing number of publications appearing in academic journals. For example, Magnan (2006) reported that 6.8% of research articles published in *The Modern Language Journal* from 1995 to 2005 were MMR.

There are two additional reasons that contribute to this turn toward MMR research in LTA. First, the definition of the key construct of interest in LTA, language ability, has embraced multicomponential communicative competence models, prompting researchers to seek to understand the linguistic, psychosocial, political, and cultural dimensions of language competence and its discursive interactions among the dimensions (Abbuhl & Mackey, 2008). Accordingly, MMR researchers seek to understand the complex nature of language acquisition and interaction among various language users in specific sociopolitical and cultural contexts.

Second, these research endeavors indicate LTA and educational researchers' recognition of MMR not only as a legitimate inquiry approach, but also as a necessity for establishing strong validity frameworks to warrant inquiry claims. Historically, the notion of validity was conceptualized as a set of scientific and objective criteria used to determine the degree to which the findings of scientific investigations could approximate the truth with certainty. Therefore, the notion of test validity was equated with objectivity and accuracy in measuring unobservable latent abilities, such as language proficiency. This early view was clearly postulated in Guilford's (1946) definition of validity that a test is valid for anything with which it correlates. A more common definition of validity at that time was that a test is valid to the extent that it measures what it is supposed to measure.

Such views were largely rejected by contemporary scholars partly because these conceptualizations failed to recognize testing and assessment as social practices (Cronbach, 1988). As noted by Cronbach:

> As psychological science generates new concepts, test interpretation will have to be reconsidered. Also, because psychological and educational tests influence who gets what in society, fresh challenges follow shifts in social power or social philosophy. (p. 5)

Furthermore, with the rise of the interpretative and constructive paradigms in the 1980s (Lincoln & Guba, 1985), scholars came to question the adequacy of value-detached test validity and the existence of any foundation to be used as absolute criteria for legitimizing knowledge claims. Messick's (1989) seminal

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

article on test validity was reflective of the paradigmatic shift and expansion in that era and succinctly summed up the evolving notion of validity:

> What is to be validated is not the test or observation device as such but the inferences derived from test scores or other indicators—inferences about score meaning or interpretation and about the implications for action that the interpretation entails . . . . Validity always refers to the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores. (p. 13)

As the quoted text illustrates, Messick (1989) emphasized validity as evaluative judgments about the quality of interpretations and use based on test scores. Subsequently, scholars explored how human experiences and interpretations are context-bound, intersubjective, and co-constructed through interactions with other constituents (Schwandt & Jang, 2004). The rise of this interpretive camp led to the historic paradigm wars in the early 1990s (Gage, 1989), which prompted researchers to recognize MMR as a third wave of the methodological movement. Table 1 provides a historical overview of the expanded notion of validity organized across main inquiry traditions.

The issue and conceptualization of validity in MMR is a topic of recent interest and discussion among researchers (e.g., Creswell & Plano Clark, 2007; Dellinger & Leech, 2007; Onwuegbuzie & Johnson, 2006). Publications in this area have emerged in response to the concern that evaluating validity claims in mixed methods investigations are particularly complex because of the combination of the "complementary strengths and non-overlapping weaknesses of quantitative and qualitative research" (Onwuegbuzie & Johnson, 2006, p. 48). Essentially, there is recognition that validity is conceptualized differently in quantitative and qualitative paradigms, as such when these methods are mixed, regardless of the purpose of mixing, a complementary notion of validity needs to be articulated. Teddlie and Tashakkori (2003) emphasized inference quality as a key validity concept that includes conceptual consistency, interpretive agreement, and interpretive distinctiveness. Greene (2011) argued that inference quality should "embrace argument alongside evidence, divergence alongside agreement, contextual understanding alongside causal explanations, and inclusion of the uses and action consequences of the inferences rendered" (p. 89). MMR researchers seek to gain ecological and authentic understandings of cultural contexts (Kirkhart, 2005) beyond causal claims based on the observation of samples drawn from an unknown population.

Specifically in the field of LTA, researchers seek evidence supporting the validity of assessment practice by integrating the nuanced understanding of human experiences and relations in LTA contexts into the traditional validity discourse. LTA researchers raised moral and ethical issues arising from the impact of testing on the lives of people and society and questioned if intention-based construct validity claims obscure social, consequential responsibilities of test developers and researchers (see the special issue *Ethics in Language Testing*; Davies, 1997).

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

Table 1. Evolving Notion of Validity

	Measurement	Quantitative Inquiry	Qualitative Inquiry
1940s	• A way to ensure objectivity and accuracy in measuring unobservable human traits • A test is valid for anything with which it correlates (Guilford, 1946)	A set of scientific and objective criteria for determining the degree to which research findings can approximate the truth with certainty supported by empirical evidence	
1950s	1954 Standards: Four types of validity—content, concurrent, predictive & construct validity	Campbell & Fiske (1959): Convergent and discriminant validity (MTMM)	
1960s–70s	• Cronbach & Meehl (1955): Construct validity • 1966 & 1974 Standards: Three aspects of validity, criterion-related, content, and construct, called the "holy trinity of validity" (Guion, 1980)	Campbell & Stanley (1966) and Cook & Campbell (1979): Threats to internal and external validity in randomized controlled design (a golden standard)	Denzin (1978): Triangulation of data, methods, investigators, and theories
1980s	• 1985 Standards: Appropriateness, meaningfulness, and usefulness of inferences from test scores and validation as the process of accumulating evidence • Messick (1989): A unitary view of construct validity—value implications and social consequences	Cronbach (1988): The logic of evaluative argument from functional, political, operational, economic, and explanatory perspectives	• Miles & Huberman (1984): Internal and external validity • Lincoln & Guba (1985): Trustworthiness-credibility (internal), transferability (external), dependability (reliability), confirmability (objectivity)
1990s	• Kane (1992), Shepard (1992): Validity as arguments based on evidence for and against proposed interpretations and use • Moss (1994): Can there be validity without reliability? • 1999 Standards: Construct validity as an overarching concept and emphasis on consequential evidence	Internal and external validity for randomized control trials (RCT)	Maxwell (1992): Five types of validity
2000s	• Kane (2002, 2006): Validity arguments • Bachman (2005): Assessment-use argument	**Mixed methods inquiry** • Teddlie & Tashakkori (2003): Inference quality that combines design quality and interpretive rigor • Dellinger & Leech (2007): Construct validation as a MM validation framework • Onwuegbuzie & Johnson (2006): Legitimacy as an overarching validity term • Greene (2008): "Criteria for warranting inferences need to be blended or created anew" (p. 15)	

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

It is from this historical vantage point that we intend to examine MMR in LTA by analyzing empirical research in the field published in the last 5 years. Accordingly, we begin with a general discussion of the conceptualizations of MMR by focusing on some of its key tenets as well as its various purposes and designs. We continue with an examination of the MMR research in LTA. Based on an evaluation of emerging trends of MMR in LTA, we then offer some general suggestions for how to advance MMR in LTA. These suggestions are not solely based on our belief in the value of MMR, but our contention that the contribution of MMR is integral to researching the complexities of the language construct, which necessitates integration of multiple approaches through multiple perspectives. Our discussion now continues with an overview of MMR not simply as a method, but as an alternative paradigm to traditionally qualitative and quantitative inquiry.

MMR AS AN ALTERNATIVE PARADIGM

As we noted earlier, a common definition of MMR is an inquiry approach that mixes both qualitative and quantitative methods within a single study (Greene, 2007; Johnson et al., 2007; Tashakkori & Teddlie, 2003). This relatively simple definition of MMR lends itself to varying interpretations. As such, there are inconsistencies in the way that MMR is conceptualized despite its increasing applications in LTA and education with part of the difference arising from its literal meaning (i.e., mixing methods). In particular, a main conceptual inconsistency is reflected in the view that MMR constitutes a methodical mixing; that is, it combines qualitative with quantitative methods. While this view contributed to expanding research techniques, it fails to address intricate issues associated with whether or not different methods are tied to particular research paradigms. Patton (2002) notes:

> My practical (and controversial) view is that one can learn to be a good interviewer or observer, and learn to make sense of the resulting data, without first engaging in deep epistemological reflection and philosophical study. . . . [One can] simply conduct interviews and gather observational data to answer concrete program and organizational questions without working explicitly with a particular theoretical, paradigmatic, or philosophical perspectives . . . without making a paradigmatic or philosophical pledge of allegiance. (pp. 69, 145)

Such a methodical view is criticized for its emphasis on technical mixing, resulting from a lack of attention to the logic of justification that can help researchers question the nature of social realities, clarify the relationship between the researcher and the researched, and bring forward the epistemological assumptions made when generating knowledge[2] claims.

Conversely, the view of MMR as a methodology posits it as an alternative paradigm that is intended to contribute to overcoming the existing paradigmatic divide. This perspective posits that paradigmatic differences are reconcilable

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

through a new emergent paradigm that actively promotes the mixing of methods across research traditions as Greene (2007) notes:

> Mixed method inquiry is an approach to investigate the social world that ideally involves more than one methodological tradition and thus more than one way of knowing, along with more than one kind of technique for gathering, analyzing, and representing human phenomena, all for the purpose of better understanding. (p. 119)

This view encourages MMR researchers to engage in philosophical discussions about the nature of being (ontological), different forms of knowledge (epistemological), and particular ways of knowing (methodological). At the methodological level, mixing can take place throughout the entire inquiry process including one or more of the following aspects:

- Inquiry problems with qualitative and quantitative inquiry approaches
- Sampling involving probability and purposeful procedures
- Data with numerical and textual types
- Data analysis involving inferential statistics and interpretive phenomenological analysis
- Inductive and deductive reasoning for theory building and testing
- Interpretations based on emic and etic representations
- Focus on particularity and generality of an inquiry phenomenon

In this article, we conceptualize MMR as a methodology situated within a specific paradigm of inquiry rather than methodical techniques and seek to understand the trends of MMR in the field of LTA.

KEY TENETS OF MMR

As an alternative paradigm, MMR, endorses pluralism in that the complex nature of a social world is best understood from multiple perspectives and multiple methodological lenses each offering partial, yet valuable, insights. MMR researchers may adopt various stances including: pragmatism, dialectical and transformative. Pragmatists take the view that philosophical assumptions are logically independent and can be mixed and matched to achieve the most appropriate answer to a given research problem (Johnson et al., 2007; Morgan, 2007; Reichardt & Cook, 1979). Rejecting a traditional dualistic divide, they prioritize practical problems crossing paradigmatic boundaries. The pragmatists choose methods that are most likely to provide data useful for answering research questions given a research context and available resources and informing the decision-making processes that are practically useful and contextually relevant (Datta, 1997; Johnson & Onwuegbuzie, 2004).

Researchers adopting a dialectical stance (Greene, 2007; Greene & Caracelli, 1997) recognize paradigms as social constructions that are historically bound in cultural contexts. MMR is viewed as a way of intellectually engaging

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

contradictory or hidden views, values, and assumptions by juxtaposing them with contesting arguments. In this way, the dialectical researchers seek to gain critical knowledge claims from different paradigms or different value positions.

The transformative stance (Mertens, 2007, 2010) acknowledges inequalities and power differentials that exist in society and views MMR as a way to bring forward the voices of socially marginalized people. Transformative researchers (Mertens, 2007) believe that MMR can provide a basis for social change by seeking input from those who were traditionally excluded from power positions.

Although there are differences among these three stances, there are also commonalities. Namely, all the MMR stances posit that social inquiries can have multiple purposes, such as theory building and testing for a better understanding, improving programs and informing decision-making processes, and promoting social justice and equity. They further agree that human and social phenomena are too complicated and multilayered to be known through a single inquiry lens. By adopting an MMR methodological stance, researchers are interested in actively engaging both particular local details and underlying patterns and generality. They strive to gain more comprehensive, insightful, and richer understandings that are more defensible and credible (Cook, 1985). Understanding these stances provides a lens through which we can make explicit the underlying purposes and designs of MMR.

TYPES OF MMR BY PURPOSE AND DESIGN

MMR cannot be wholly understood through a discussion of its underlying stances; as with any other approach, MMR demands a clear rationale for its purpose and design (Maxwell & Loomis, 2003). There are four broad purposes of MMR, including triangulation, complementarity, development, and initiation. Each of these purposes has identifying characteristics at various stages of inquiry, from formulating research questions and collecting and analyzing data to interpreting and synthesizing findings.

Triangulation is one of the most common purposes of MMR and is used to confirm and cross-validate findings while offsetting biases associated with different methods (Denzin, 1978; Mathison, 1988). Originating from Campbell and Fiske's multitrait-multimethod matrix (1959), the notion of triangulation has been expanded to entail triangulations of data, investigators, theory, and methodology (Denzin, 1978). MMR mainly concerns the methodological triangulation in that it involves independent data collection and analysis and joint interpretations at a synthesis stage. Parallel comparison for triangulation requires that different methods elicit the same phenomenon, samples used for different methods are independent, and each method should be given an equal weight. Although convergence is the primary purpose of triangulation, researchers acknowledge that MMR cannot only provide converging but also inconsistent and even contradictory results, which prompt researchers to seek alternative explanations for the findings (Mathison, 1988).

Another common purpose of MMR is complementarity in that it seeks to elaborate, clarify, and explain the results from one method with the results

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

from another method. It may address overlapping or different yet interrelated aspects of the same phenomenon to seek the breadth of the inquiry (e.g., teacher beliefs and practice or program process and outcomes). This latter purpose of complementarity is referred to as an expansion model of MMR (Greene, Caracelli, & Graham, 1989). Data from different methods may be collected concurrently and analyzed independently (Teddlie & Tashakkori, 2006). Alternatively, the complementary MMR can be used in other MMR designs, such as triangulation, developmental, and nested designs.

The nested MMR design typically places a lower priority method within a dominant one. Data from the lower priority method are used to elaborate, illustrate, and enrich the findings from the dominant method. For example, an experimental study may be conducted in which qualitative methods are embedded, or it may be a largely qualitative inquiry with a complementary quantitative method.

A third, and equally common, purpose of MMR investigations is developmental MMR. In a developmental MMR design, data from one method are used to guide subsequent data collection and interpretations or to complement and explain results from the preceding method. The full cycle of the developmental MMR features abductive reasoning (Fulcher, 1996; Mislevy, 1995) involving (a) inductive case observations used to develop a working theory; (b) hypothetico-deductive testing used to test the theory for its generality; and (c) the application of theory back to cases to confirm or disconfirm its generality. The decision of how to sequence methods may depend on the availability of relevant theoretical frameworks. MMR can start with an exploratory qualitative inquiry for inductive case observations and test the emerging theoretical framework. Creswell (2007) and others call this MMR design an *exploratory sequential*.

Alternatively, an explanatory sequential design is used to test an existing theory for its construct congruence and generalizability to the given target population, and subsequently, the core characteristics of the theory, with specific cases in hand, are examined. Note that the developmental MMR does not need to be linear nor complete in cycle. The process is iterative involving both inductive and deductive reasoning processes (Jang, 2013).

A fourth purpose of MMR, initiation, seeks to discover any potential paradox and contradictions underlying social inquiry problems (Greene et al., 1989). Researchers may mix methods purposefully to recast questions based on results from one method and analyze both consistencies and discrepancies between qualitative and quantitative findings for fresh insights.

The various MMR designs we have reviewed so far are useful for clarifying its purposes; however, in practice, MMR is too complex to be adequately represented by any single design (Maxwell & Loomis, 2003; Teddlie & Tashakkori, 2006). MMR designs should not be applied rigidly or exclusively (Tashakkori & Teddlie, 2010); they can be further mixed or modified to create a new component of the design.

One important area for growth in MMR is data integration and synthesis. Parallel presentations of findings by method or question type limit the synergistic effects of MMR, prompting the reader to question its intent. This lack of integration in MMR can be attributable to logistical constraints, such as limited

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

space in most journals to publishing the entire MMR investigation or pressure to amass large numbers of publications in academia. Fully integrated designs mix qualitative and quantitative approaches in an interactive and iterative manner throughout a study. Preliminary results from one method may be transformed to a form compatible with the other for further analysis or be consolidated to create blended variables to further examine new insights. Therefore, mixing methods does not automatically guarantee stronger claims; however, fully integrated MMR has the potential to provide a holistic understanding. We now turn to a discussion of our review of MMR in LTA. We begin with a discussion of our process for identifying these publications.

THE REVIEW OF MMR IN LTA

Our review process of MMR in LTA involved several collaborative stages among the authors in order to address three main purposes: (a) identifying all of the peer-reviewed journal articles employing MMR in LTA for the past 5 years; (b) reading and analyzing each article to understand the rationale for and the approaches to MMR; and (c) synthesizing the analyses in order to highlight some key trends of MMR. We now discuss each of these stages in detail.

Identifying MMR Studies

The process of reviewing the literature involved identifying and applying specific inclusion and exclusion criteria for studies that employed MMR published between January 2007 and July 2013. These criteria included the fact that the research needed to be peer reviewed and published in an English-language journal addressing LTA for learners of any age. Additionally, we identified key terms to use to search research databases and specific journals. We excluded any study that either did not meet our specified criteria or did not report on mixed methods data even if it was mentioned as a method of data collection. We also excluded studies that exclusively used multiple quantitative or qualitative data, but included studies that used both qualitative and quantitative data. See Appendix A for the summary of criteria used for identifying MMR studies. We identified 32 empirical studies that met the criteria for MMR that we subsequently analyzed.

Analyzing MMR Examples

Upon identifying an initial set of research studies, we worked collaboratively to determine specific areas of foci to use for the analysis of each study. Initially, each researcher analyzed two studies independently, and then we met to discuss the findings as well as the categories that we used to analyze the articles. Through a collaborative process we revised these categories and agreed on a final list of the areas of foci to use to analyze the language assessment research employing mixed methods. Specifically, we used the following categories: population of interest, study participants, language modality, research context, research questions and/or purpose(s), rationale for the use of mixed

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

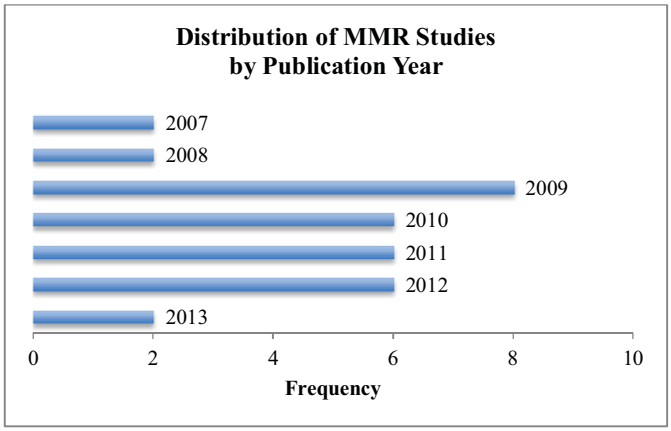


Figure 1. Distribution of MMR studies by publication year.

methods, timing of mixing, evidence of collaboration, data sources, data analysis approaches, presentation of results, and synergetic effects (see Appendix B). We further examined ways in which the studies addressed validity issues. Additionally, we each recorded additional comments highlighting important facets of each article. We each analyzed a fraction of the articles. Upon completion of this work, we met to discuss the results in depth and assess if all of the questions had been fully answered. In some cases, we revisited the research articles to ensure that our analyses were fully completed.

Synthesizing the Analyses

In the final stage of the review process, we met once again to synthesize the findings by analyzing trends and discussing our observations and emergent issues. We synthesized the findings by discussing each of the categories of analysis independently and also by examining overall trends of MMR.

TRENDS OF MMR IN LTA

Of 27 journals that were reviewed, 32 studies from 15 journals satisfied the specific inclusion criteria. Overall, 21 studies across various educational sectors focused mainly on learners as study participants. Eleven studies involved more than one participant type (e.g., raters). Teachers were recruited as raters in nine studies. There were two studies that involved administrators as well as students and teachers.

Figures 1–3 summarize the distribution of the 32 MMR studies according to publication year, language component, and study population.

As shown in the figures, MMR publications have increased over the past 10 years (note that the data for 2013 are incomplete as we reviewed articles

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

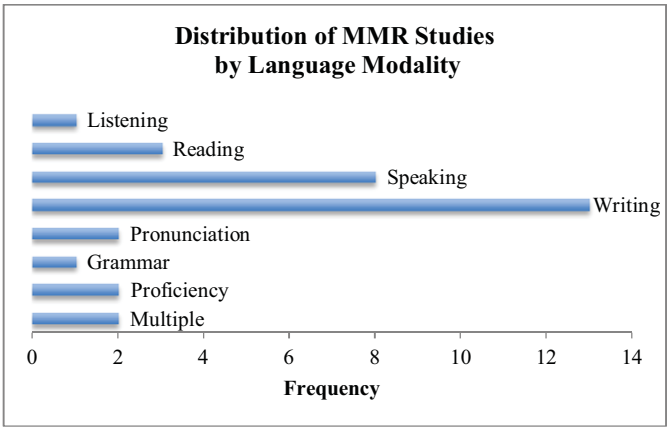


Figure 2. Distribution of MMR studies by language modality.

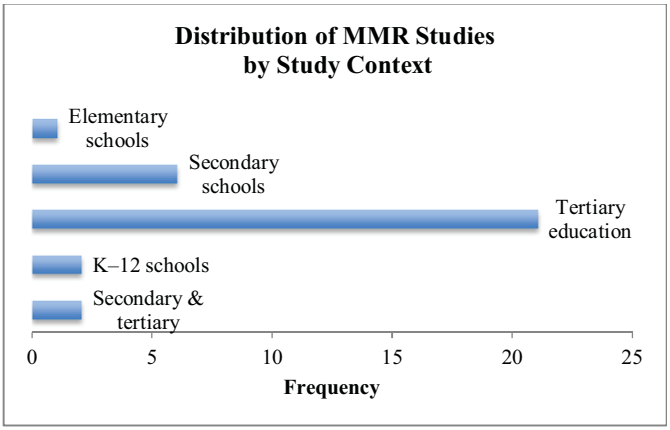


Figure 3. Distribution of MMR studies by study population.

published up to July 2013). These MMR investigations most frequently focused on writing assessment as compared to other language skills. The study contexts varied from elementary to postsecondary school sectors while the latter tended to be dominant, which is consistent with the fact that large-scale language tests are predominantly used to assess adult learners rather than young learners.

Purposes of Mixing

Before we delve into our analysis of the purposes of mixing among the empirical studies we identified, it is necessary to highlight the fact that not all of these studies explicitly acknowledged or defined their investigation as MMR, but rather, it was through our reading and analysis of the articles that we identified them as MMR. The interested reader might wonder how we uncovered

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

Table 2. Distribution of MMR According to Purposes of Mixing

Purpose	Study Examples
Triangulation	Baker (2012); Isaacs (2008); Jang (2009); Jang & Roussos (2009)**; Lee & Greene (2007); Nakatsuhara (2011)*; Perrone (2011)**; Plakans & Gebril (2012); Yu (2010)**
Complementarity expansion	Anthony (2009)*; Baker (2010)*; Barkaoui (2010)*; Barkaoui (2011)*; Colby-Kelly & Turner (2007)*; Derwing, Munro, & Thomson (2008)*; Hamid et al. (2009); Harsch & Martin (2012)*; Hyland (2009)*; Kim (2009); Knoch (2009); Lee & Winke (2013); Perrone (2011); Poonpon (2011); Turner (2009); Uchikoshi & Maniates (2010)*; Winke (2011); Xi (2010); Yu (2010)**; Wiseman (2012)*; Yin, Sims, & Cothran (2012)*
Development	Ghanbari et al. (2012)*; Isaacs & Trofimovich (2012); Jang & Roussos (2009)**; Kim & Jang (2009)*; Jang (2009)**; Baker (2010)
Initiation	DiPardo et al. (2011)
Transformation	Hamid et al. (2009)**

Note: *Did not explicitly state the purpose for employing a mixed methods study. The classification is based on our analysis.
**Study is placed in multiple categories.

these studies when there was no explicit mention of MMR; the credit is owed to powerful search engines that search for key terms not only in the body of published articles, but also in reference lists. Oftentimes, the studies that did not explicitly address MMR in the body of their articles cited other publications with MMR explicitly stated in their titles (and hence searchable through the reference lists). It is also possible that additional MMR investigations have been published in the last five years that we were not able to uncover, and hence did not include in this review, because they did not explicitly acknowledge mixing.

Among the 32 MMR studies we reviewed, we identified several studies that mixed methods for multiple purposes, which made it difficult for us to categorize them using the MMR typologies (Creswell & Plano Clark, 2007; Greene et al., 1989) discussed earlier. We acknowledged this multiplicity in a summary of our analyses presented in Table 2. Note that Table 2 also includes two additional MMR typologies (initiation and transformative) that reflect researchers' paradigmatic orientations to MMR.

Overall, our review showed that complementary mixing was the most frequently identified as the purpose of mixing. We identified 18 studies that used mixed methods to elaborate, clarify, and illustrate the results. There were 12 studies that combined methods for triangulation purposes to determine if there was convergence or corroboration of the results emerging from the different methods employed. The prevalent use of MMR for the triangulation and complementarity purposes in LTA is consistent with the application of MMR to applied linguistics and second language (L2) acquisition (Hashemi, 2012).

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

One common concern regarding the application of MMR for triangulation arises when mixed data point to divergence and inconsistency. Researchers express concerns when they encounter inconsistent or even contradictory findings from mixed methods data. Unlike the traditional notion of triangulation used for mixing multiple methods without crossing the methodological border (e.g., qualitative interviews with open-ended observations), the purpose of triangulation used in MMR should not solely focus on producing converging evidence to support a singularly true proposition (Mathison, 1988). MMR scholars may choose to welcome not only converging but also inconsistent or even contradictory findings as plausible explanations that need to be thoroughly validated. This welcoming of contradiction was exemplified in Yu's (2010) MMR investigation.

In Yu's study (2010), 157 students read extended English texts, presented on computer or in print, and then wrote summaries on paper in English and Chinese. The study also used questionnaires and interviews to gather data about the students' perceptions about presentation mode and computer familiarity. The students' summary task performance data were analyzed using univariate analyses to determine the effect of presentation mode and computer familiarity. The questionnaire data were factor analyzed to identify factors associated with computer familiarity. The interview data were thematically analyzed to understand psychological differences in students' experience with different presentation modes. Although the analysis of each data source was performed and presented separately, Yu discussed both converging and diverging results in the discussion. He stated that although the presentation modes did not affect students' summarization task performance, except for a main effect on the length of Chinese summaries (converging evidence), the results from the qualitative data suggested physical and psychological differences in students' experience with the two presentation modes and their potential effect on the task performance.

Five studies were identified that mixed methods for developmental purposes. In these studies, findings from earlier phases of research were used to inform subsequent phases. The developmental purpose of MMR is often confusing because its sequential order of data collection is necessitated by the interdependency of mixed methods. Simply collecting one data type, say using a questionnaire, before another type (e.g., interviews) does not make a study developmental. It is considered developmental if the results from the first data source are used to guide a facet of the second. For example, questionnaire data could be used to inform interview protocols or determine criteria for interview sampling.

Ghanbari, Barati, and Moinzadeh's (2012) examination of Iranian English as a foreign language (EFL) raters' use and perceptions of a writing rating scale exemplifies a developmental MMR design. They interviewed 10 expert raters and used the results from the interviews to develop a questionnaire. They interviewed 7 (out of 30 questionnaire participants) to seek further information about their responses to the questionnaire. Another developmental MMR example is illustrated by Isaacs and Trofimovich's (2012) investigation. They designed a sequential MMR over three phases. In phase 1, 60 listeners rated 40 Francophone English language learners' speech samples. In phase 2, the same speech samples were presented to teachers to seek their written reports on the

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

aspects of speech they attended to. In the final phase, the researchers identified core speech measures that distinguished among L2 comprehensibility levels to develop a guideline for L2 comprehensibility scales.

We identified only one study (DiPardo, Storms, & Selland, 2011) that mixed methods for initiation purposes in order to uncover contradictions or reveal new perspectives. There was also only one investigation by Hamid, Sussex, and Khan (2009) among the 32 we analyzed that mixed methods for transformative purposes to address inequities and advocate social change. Specifically, Hamid et al. examined the private tutoring in English (PT-E) practices in a disadvantaged rural area of Bangladesh. They insisted that although PT-E is considered as a necessary condition for academic success in Bangladesh, the quantity and quality of the participation are heavily mediated by the family affordability in the studied area, where mainstream schooling and the self-study resources are shown to be inadequate. We now turn our attention to the study contexts.

Context

The majority of the studies took place predominantly in higher education learning contexts with adult immigrant or university populations. For example, Hyland (2009) involved first-year college students who took a school-mandated academic writing assessment to examine referencing practices of students who speak English as a first language (EL1) and learners who speak English as a second language (EL2). Nine studies focused on the K–12 context in general, six of which addressed issues related to secondary adolescent learners. It is notable that there were only a few studies that focused primarily on young language learners. Uchikoshi and Maniates's (2010) MMR research examined bilingual literacy development involving Spanish-English and Cantonese-English bilingual children to investigate the bilingual classroom practices and home–school connections. Nakatsuhara's (2011) investigation also placed a focus on younger learners, specifically on upper secondary EFL learners in Japan. She investigated the nature of co-constructed interactions in group-speaking tests.

Collaboration

To what extent are the MMR investigations in these contexts a result of collaborative endeavors? We chose collaboration as a category of analysis in this review because of its importance to MMR. Most researchers have specialized in their academic disciplines and professional training in either a predominantly qualitative or quantitative tradition, the so-called trained incapacities (Reiss, 1968, cited in Bryman, 2007). Dörnyei (2007) posits that this localized expertise may be a barrier to conducting MMR. One common solution to this challenge is for researchers to bring together teams with a range of expertise. Interestingly, there was little mention of collaboration in the MMR studies we analyzed beyond the fact that the reviewed manuscripts were sometimes authored by multiple researchers or made references to research assistants coding data mainly to establish reliability estimates. For example, Perrone's (2011) research engaged additional researchers to videotape and also code data to enhance reliability.

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

No explicit mention of collaborative teamwork in MMR may suggest that LTA, compared to other disciplines, is equipped with researchers who specialize in both research traditions. If findings from mixed data are well integrated to generate insights that data from one method alone cannot achieve, then, the trained capacity in MMR is much more desirable than developing a team of researchers, each of whom has a strong inclination or training in either tradition. However, Hesse-Biber and Leavy (2006) question the value of MMR done by researchers inadequately trained in both inquiry approaches. Bryman (2007) warns that "the presence of skill specialisms may lead to compartmentalization of roles and responsibilities that can hinder the integration of findings" (p. 16). Our observation that there was no explicit mention of collaborative teamwork for MMR is tentative and needs to be further probed by looking into the current trend of training in LTA in graduate programs. As Dörnyei (2007) asks, "Apart from a relatively small number of unique, methodologically ambidextrous specimen, can we assume the vision of a multimethodologically savvy new breed of researchers is realistic?" (p. 46). We now turn our attention to the sources of data used in the MMR studies.

Data Sources

Together, the MMR studies incorporated a range of data sources that primarily included test data, ratings (e.g., scores from oral interviews, writing scripts), written accounts (e.g., reflective essays), interviews, questionnaires, and think-aloud protocols. To a lesser extent, classroom observations and field notes were also used as data sources. Among the 32 studies reviewed, 13 focused on writing assessment and another 8 focused on issues related to speaking assessment. Three studies examined reading, while there were four studies that examined all four modalities (see Figure 2).

Quantitative data sources predominantly originated from ratings of students' written or spoken responses, questionnaires, and scored responses to proficiency or achievement tests. Considering that 13 out of the 32 reviewed studies focused on writing assessment, it is not surprising that the most frequently observed quantitative data source was derived from students' written responses, which were subsequently treated to generate mixed data, rating scores, and written accounts. This treatment exemplifies the case of mixing at the stage of analysis by quantizing qualitative data. Some studies (e.g., Poonpon, 2011) also analyzed students' written and spoken responses qualitatively and quantitatively simultaneously in order to triangulate the findings to offer confirmatory evidence supporting the accuracy of rating in speaking and writing assessment.

Other notable mixing of data was observed between (a) students' test responses and their verbal reports from either think-alouds or interviews, (b) raters' ratings and their verbal/written accounts, or (c) quantitative data from closed or Likert-scale responses to questionnaires and written responses from open-ended questions included in the same questionnaire. These pairs of mixing were often used for triangulation and were also predominantly used to complement each other while qualitative data played a complementary role by illustrating proficiency levels inferred from the scores with specific examples. For

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

example, Winke (2011) investigated the "perceived effectiveness" (p. 628) of the English Language Proficiency Assessment (ELPA) by educators in order to investigate validity claims. The primary data source was a set of responses to a survey that comprised (a) items gathering participants' demographic information; (b) belief statements with Likert-scale response options, as well as space for written comments; and (c) five open-ended questions. Winke reported that the quantitative data were used to identify response patterns, and that the qualitative data were used for illustrative purposes.

As we discussed earlier, MMR cannot be reduced to the inclusion of both qualitative and quantitative data included in a single study; there is a need for integration. The next section of our analysis addresses the MMR studies' approaches to data analysis and integration.

Data Analysis and Integration

Integrative data analysis and interpretation of mixed methods are critical in MMR. Our review showed that some studies did not provide sufficient information about their data analysis, and that the majority of the studies ($n = 23$) reviewed kept both analysis and interpretation of the two data types separate. There were nine studies that analyzed and interpreted integratively.

Studies that kept the two types of data separate presented results according to data type. For example, Baker (2010) compared raters' scores between high- and low-stake testing conditions in order to examine differences in grading in the raters' perceptions. Results from each data source and phase were reported separately and integrated in the discussion of all the study findings in the later sections of the article.

Similarly, some studies analyzed mixed methods data separately to answer different research questions. In such cases, results were presented by questions (not through a quantitative or qualitative distinction), and integration took place at the discussion stage. For example, Kim (2009) analyzed 80 speech samples using the Rasch model and independently analyzed teachers' written comments thematically to answer two separate questions respectively. She brought the results together in the discussion by highlighting that "the inclusion of a qualitative analysis provided insight into the different ways in which NS and NNS teachers assessed students' oral language performance, above and beyond findings from the quantitative analysis alone" (p. 210).

We also examined the MMR data analysis and integration of the studies by examining them according to established MMR data analysis strategies. Caracelli and Greene (1993) discuss four integrative mixed methods data analytic strategies: data transformation, typology development, extreme case analysis, and data consolidation. Data transformation involves the conversion of one data type into the other (e.g., quantitizing qualitative data by assigning numerical or categorical codes or qualitizing quantitative data into a narrative to be compared with thematically analyzed qualitative data). Typology development is used when one type of data yields typology (a set of distinct categories that explain dimensions and patterns) and the resulting typology is used as an explanatory variable for subsequent analysis. Extreme case analysis is used to

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

confirm or disconfirm findings from one data type by using cases from the other data type. Data consolidation is used through a joint analysis of both data types to create consolidated data sets for further analyses.

None of the studies we reviewed used typology development or data consolidation strategies. A few studies we reviewed analyzed data through integrative mixed methods data analytic strategies including data transformation and extreme case analysis. The most frequent data transformation was done by coding students' writing and speech samples both inductively and deductively. This integrative analysis allowed for a direct comparison of the results and highlighted both overlapping and non-overlapping aspects of the phenomenon under investigation. For example, Nakatsuhara (2011) analyzed group oral tests from 269 Japanese upper secondary school students in order to investigate the influences of students' extraversion levels and oral proficiency levels on their conversational styles in different-sized groups. The data sources included several questionnaires, students' oral proficiency levels as measured using a rating scale, and students' performances on group oral tasks. Students' oral performance data were used to quantify conversational styles in multiple ways, and these results were used to conduct several multiple regression analyses. However, Nakatsuhara also analyzed students' performances qualitatively "to illuminate the quantitative results" (p. 491). Nakatsuhara explained the integrated data analyses approach:

> . . . the coded data were analysed through multiple regression to provide a generalizable picture of the data. However, CA [conversational analysis] is generally hostile towards explicit coding and quanti-fication (Schegloff, 1993). Therefore, when applying CA approaches in the second part of the results, the analysis focuses on interpretation, avoiding the use of explicit quantita-tive language. The two perspectives are seen as complementary, the CA providing insights that help to explain the statistical results. (p. 492)

Extreme case analysis is an integrative analytic strategy used to confirm or disconfirm findings from one data source using cases from the other type (Caracelli & Greene, 1993). There were two studies that used additional case analyses (Jang, 2009; Lee & Greene, 2007). In one of these studies, Lee and Greene (2007) investigated the relationships between graduate English as a second language (ESL) students' placement test scores and their academic performance measures (i.e., grade point average [GPA], faculty evaluations, and student self-assessment). Using correlational analyses, they triangulated students' perspectives with that of the faculty by mixing data from questionnaires with interviews. They further analyzed mixed data using extreme case analysis and joint display strategies to shed light on inconsistencies and contradictions across the data sets. They identified three lower-level cases and one higher-level case based on statistical analysis that detected outliers (two standard deviations below and above the placement and GPA means). Through in-depth case analyses, Lee and Greene sought to explain the case students' perspectives about academic

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

success. The joint display of the quantitative and qualitative results by participant type (students vs. faculty) supported the predictive validity of the placement measure and further contributed to revealing the nuanced nature of academic learning and achievement.

Synergy refers to the added value of MMR as a result of integration of the findings from mixed methods data analysis. This synergistic effect is greater than if only qualitative or only quantitative approaches are taken (Nastasi, Hitchcock, & Brown, 2010). Overall, the degree of integration in the studies we reviewed is limited because the majority of the studies kept the qualitative and quantitative components separate rather than integrative. A few studies that we regarded as having synergism are among those using the most complex MMR designs (e.g., Jang, 2009; Nastasi et al., 2010).

One of the core principles of synergism is the dialectical stance that we introduced earlier and can be exemplified in Jang's (2009) research, which took this stance to provoke potential issues with cognitive diagnostic modeling (CDM) by using case analyses. Jang was interested in the extent to which reading skills on an assessment could be used for diagnostic purposes. Her study generated diagnostic feedback using CDM and sought out students' reaction to it. Students' responses to the usefulness of the (quantitative) diagnostic feedback contradicted Jang's observations of the teachers' use of the feedback. She observed that teachers did not have time to address all the skills, prompting her to question the validity of students' reports. Consequently, she sought evidence for counterarguments using additional methods. The attempts to reconcile these differences highlight the tenets of the dialectical stance and further demonstrate the synergistic effects prompted by the use of multiple methods.

Clearly, MMR does not guarantee a sum greater than its parts (Dörnyei, 2007; Hesse-Biber & Leavy, 2006). As we have discussed, MMR necessitates the integration of quantitative and qualitative data at some stage of the research investigation. Many examples of MMR demonstrate integration of data at the discussion stage after analyzing and presenting each type of data separately (e.g., Barkaoui, 2011; Kim, 2009). Our review suggests that the application of MMR in LTA has not reached its fullest potential due to a lack of integration of mixed methods data throughout the inquiry procedure, which seems to be a common trend of MMR across disciplines.

Dörnyei (2007) discusses three reasons why people do not mix more actively and pinpoints practical constraints (such as page limits in journal publication) or pressures for more publications:

> Finally, there may be a third reason for the paucity of mixed methods studies. The publication pressures of the current research climate promote piecemeal publications in general and therefore even when a project involves methods mixing, the authors might choose to try and publish the results of the different phases separately. (p. 175)

More examples of successful practice employing integrative data analysis strategies in mixed method contexts are needed.

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

Role of MMR in Expanding the Validity Framework

We introduced this article with a discussion of the need for MMR being partly prompted by LTA researchers' desire to expand validity frameworks. It is fitting that we now revisit this issue following our analysis of 32 MMR examples. As illustrated in Table 1, the notion of validity has historically evolved throughout the paradigm shifts in the fields of education and assessment. We carefully examined the MMR studies by focusing on if and how the authors of the investigations addressed or conceptualized validity.

The results of this analysis indicate that of the 32 studies included in this review, 13 did not focus on validity issues in their studies. The remaining investigations all address various facets of validity including inference quality, consequential validity, and construct validity. Due to the (relatively) high number of studies within this review that focused on raters and ratings, it is not surprising that that there were numerous instances of scoring validity (Weir, 2005) observed. For example, Harsch and Martin's (2012) mixed methods investigation discussed the development and validation process of a local rating scale based on the Common European Framework of Reference for Languages proficiency scales. As they discuss, part of the process of scale validation is piloting it for a specific purpose in its context of use. Accordingly, they involved 13 raters to use the scale to assess writing scripts (following a rigorous series of training workshops and practice sessions). Harsch and Martin developed an index for rater agreement, which they used as an indicator of scoring validity.

Table 3 presents the aspects of validity addressed by each MMR study in this review, and draws attention to the validity frameworks spanning the various studies across contexts, participants, and research foci. One example of the use of mixed methods to address this expanding framework is illustrated in Anthony's (2009) validity inquiry into the use of reflective, time-essay responses to "elicit memories of meaningful experiences" (p. 195) in undergraduate students' education. He further wished to investigate whether these experiences would reflect students' computer-mediated classroom experiences. Anthony clearly articulates the need for MMR in this study:

> The first lens I used in looking at the responses to the test was a quantitative one designed to uncover and to attempt to discount plausible rival hypotheses for my interpretations of those responses. But when the epistemological limitations of a quantitative-only perspective became evident, I applied a different, more qualitative lens. (p. 195)

As illustrated in Table 3, Anthony's investigation addressed three primary facets of validity: construct, consequential, and inference quality. This latter category, inference quality, was another facet of validity that we observed across multiple studies.

Inference quality refers to the legitimacy of the results. Such legitimacy or trust is a function of the design quality that necessitates best practices and interpretative rigor (Teddlie & Tashakkori, 2003). For example, Lee and Greene (2007) investigated the predictive validity of graduate students' English language

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

Table 3. Discussions of Validity in MMR Examples

Authors	Discussion of Validity in Study	Facet(s) of Validity Addressed
Anthony (2009)	Address the use of MMR to contribute to validity. In depth discussion of gathering evidence to support validity argumentation	Construct, consequential, inference quality
Baker (2010)	Some discussion of validation procedures for a rating scale of writing	Scoring
Baker (2012)	Validity issues not a focus of the study	N/A
Barkaoui (2010)	Validity issues not a focus of the study	N/A
Barkaoui (2011)	Validity issues briefly discussed in relation to the use of analytic or holistic scales of writing	Construct, consequential, scoring
Colby-Kelly & Turner (2007)	Adaptation of Bachman's assessment use argument framework	Consequential
Derwing et al. (2008)	Validity issues not a focus of the study	N/A
DiPardo et al. (2011)	Validity issues not a focus of the study	N/A
Ghanbari et al. (2012)	Focus on construct validity of a rating scale drawing upon an assessment use argument approach	Construct, consequential, scoring
Hamid et al. (2009)	Validity issues not (directly) a focus of the study	N/A
Harsch & Martin (2012)	Description of process of validation of the development of a local rating scale based on a preexisting language assessment framework	Construct, consequential, scoring
Hyland (2009)	Validity issues not (directly) a focus of the study	N/A
Isaacs (2008)	Empirical validation of a language measure	Ecological, construct, scoring
Isaacs & Trofimovich (2012)	Validity issues not a focus of the study	N/A
Jang (2009)	Comprehensive validity arguments; multiple sources of evidence to build validity arguments evaluating application of a diagnostic model to a reading comprehension test	Construct, consequential

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

Table 3. Continued.

Authors	Discussion of Validity in Study	Facet(s) of Validity Addressed
Jang & Roussos (2009)	Validity argument framework; focus on quality of inferences	Inference quality, construct
Kim (2009)	Address the use of MMR to contribute to validity. Discussion of validity of inferences	Inference quality, scoring
Kim & Jang (2009)	Validity of statistical analyses addressed and discussed as well as discussion of overall validity of interpretations based on the study	Inference quality, construct
Knoch (2009)	Detailed discussion of the validity process of a rating scale	
Lee & Greene (2007)	Enhancing validity of inferences through triangulation of results from various data sources	Predictive, inference quality
Lee and Winke (2013)	Briefly addressed construct validity and validity of coding categories used for analyses Implications for validity addressed in discussion	Construct
Nakatsuhara (2011)	Validity issues not a focus of the study	N/A
Perrone (2011)	Validity issues not a focus of the study	N/A
Plakans and Gebril (2012)	Address the use of MMR to contribute to validity. Other validity issues not a focus of the study	References to previous publications related to validity
Poonpon (2011)	Seeking convergence of data from multiple methods gather validity evidence	Construct, inference quality
Turner (2009)	No validity process research included as part of this study, but some discussion related to the 'valid' uses of tests for different purposes	
Uchikoshi & Maniates (2010)	Validity issues not a focus of the study	
Winke (2011)	Uses the term "broad validity" to encompass reliability, concurrent validity, predictive validity, and consequential validity to investigate test validity	Consequential

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

Table 3. Continued.

Authors	Discussion of Validity in Study	Facet(s) of Validity Addressed
Wiseman (2012)	Validity issues not a focus of the study	
Xi (2010)	Gathered evidence from manipulation of task variables to build a validity argument about a speaking task on an exam	Consequential
Yin et al. (2012)	Discussion and implications of findings for validity	Construct, consequential, inference quality
Yu (2010)	Validity issues not a focus of the study	

test scores through an MMR inquiry. In their attempt to achieve such interpretive rigor through their research design, Lee and Greene explain:

> Different methods are needed precisely to assess various facets, dimensions, and even definitions of academic performance (rather than the same facets or the same definitions, as is required in a triangulation design for convergence). This use of mixed methods to generate a more complete and elaborated understanding of the various dimensions or facets of a complex phenomenon is called complementarity. (p. 369)

The authors argue that through the use of a complementarity design they were able to incorporate a correlational design, which is typically employed in quantitative studies. They situated this design within a mixed methods investigation by gathering perspectives of multiple stakeholders and by analyzing data not only to identify consistencies but also variations and contradictions across different data sets.

Overall, our analyses of these 32 empirical investigations revealed an increasing trend toward the use of MMR in language testing and assessment. At the same time, the turn toward this alternative paradigm is not always explicitly acknowledged. The research studies reviewed clearly evidence the complexities in the purposes and uses of MMR, as well as its power to address multiple facets of LTA. At the same time, our analyses also revealed that MMR was primarily being used in higher education contexts with adult participants, and for complementary and triangulation purposes. Furthermore, the writing skill was most often the focus of the studies. It is worth noting that only a few studies that we reviewed explicitly articulate validity issues as the rationale for MMR; however, the validity issues still appear to be focused mainly on warranting claims about the test score interpretations, necessitating further investigation and development.

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

CONCLUSION

Our review of recent studies in LTA shows that the field has welcomed MMR as a legitimate inquiry approach. Because most of the studies we reviewed were not explicit about what was being mixed for what purpose and claims, it was difficult to judge what researchers had intended to achieve through mixing. For MMR to be a third wave of inquiry paradigms in LTA, we suggest that there should be more explicit discussions about the logic of justification, that is, a rationale that guides methodological decisions. Research purposes and questions should guide methodological decisions as Greene (2007) states "methodology is ever the servant of purpose, never the master" (p. 97).

We would like to see a wider range of purposes for which MMR is used beyond triangulation and complementarity. Researchers using MMR for triangulation should expect and welcome potential inconsistencies and contradictions rather than treating them as undesirable (Mathison, 1988). Ultimately, the goal of MMR is to contribute to broader, deeper, and more inclusive understandings of multilayered social worlds and human behavior. In doing so, differences should be welcomed as a potential pathway to a new understanding. To achieve these goals, it is necessary for researchers to not apply MMR designs rigidly (Tashakkori & Teddlie, 2010), but to (creatively) design methods that will wholly respond to research questions. The majority of the studies avoided mixing at the data analysis stage by keeping and reporting results separately. Although many studies integrated findings in the discussion, we encourage researchers to experiment with integrative analytic strategies for new insights. Our traditional ways of handling data and representing knowledge may need to be revisited (Greene, 2007). We observe that digital knowledge mobilization (i.e., electronic publications) allows researchers to embed links to make implicit assumptions explicit without interfering with a main flow or that researchers add a hyperlink to a video clip of participants (talking or acting) instead of translating data into numbers and text. For MMR to be a truly new paradigm, researchers shift away from quasi-mixing of quantitative and qualitative methods to abductive mixing of reasoning moving from the particular to the general, and vice versa.

We considered the validity frameworks as a guiding principle of our review of MMR research in LTA. While methodological repertoires have expanded with the increased capacity to handle various data thanks to rapid advances in computer-assisted data analytic technologies, researchers need to embrace MMR as a way of reasoning, involving making implicit assumptions explicit and seeking for greater coherence than the sum of the parts. In doing so, it is critical for researchers to consider recovering flexibility in human reasoning beyond either hypothetico-deductive or inductive reasoning alone. MMR researchers should make explicit the logic of reasoning based on practically reasoned dialectical argumentations, including not only confirming claims but also rebutting potential counterarguments. MMR should be fully grounded in specific contexts where LTA practices impact people's lives and encourage the participation of LTA users, practitioners, and policymakers. Participatory MMR practice may bring forward unheard voices in LTA practices. We conclude by welcoming further MMR that examines the complexity of power relations among various stakehold-

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

ers, which will significantly contribute to strengthening and expanding validity issues in LTA.

NOTES

1 Morgan (2007) defines paradigms as ways of experiencing the world, shared by a community of researchers (p. 53).
2 Habermas (1984) characterizes interests in knowledge according to three different purposes. First, interest in knowledge for technical purposes is aimed to predict particular outcomes. This interest is served by empirically grounded knowledge. A second kind is a practical and interpretive interest. This practical interest is served by knowledge grounded in various participants' experiences and stories. A third kind of interest takes a critical and emancipatory purpose. Intuitive knowledge is sought to help us test assumptions, rules, and deconstruct expectations that prevent us from moving forward.

ANNOTATED BIBLIOGRAPHY

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. New York, NY: Oxford University Press.

This volume thoroughly engages the reader through each major phase of research by addressing key issues, data collection, data analysis, and reporting. Each of these aforementioned sections contains discussions specific to qualitative, quantitative, and MMR, thereby providing the reader with integral information about MMR while also providing the reader with knowledge that can be used to compare and contrast the benefits and challenges to the use of each method in specific contexts. Further, the volume also includes a historical overview of the research paradigms and a summary of their strengths and weakness, along with a summary of Dörnyei's own paradigmatic stance, which highlights his call encouraging the use of MMR.

Morgan, D. L. (2007). Paradigms lost and pragmatism regained. *Journal of Mixed Methods Research*, *1*, 48–76.

Morgan introduces this article by asking a pointed question: "To what extent is combining qualitative and quantitative methods simply about how we use *methods*, as opposed to raising basic issues about the nature of research *methodology* in the social sciences?" (p. 48). In order to respond to this question, this thought-provoking article takes the reader through a historical review of the developments in research methodology in the social sciences using the concept of paradigms as a conceptual framework. Morgan presents four conceptualizations of paradigms, which include paradigms as worldviews, epistemological stances, shared beliefs in a research field, and model examples. Through a discussion of paradigms, Morgan propagates a discussion of methodological issues that arise when researchers engage in MMR. Morgan's central thesis is that a "metaphysical" paradigm in the social sciences has been exhausted and should be replaced by a "pragmatic" paradigmatic approach (p. 55).

Tashakkori, A., & Teddlie, C. (Eds.). (2010). *Handbook of mixed methods in social and behavioral research* (2nd ed.). Thousand Oaks, CA: Sage.

This handbook, weighing in at close to 900 pages, is one of the most comprehensive resources for understanding and applying MMR in any field including LTA. The handbook is divided into three sections: (a) conceptual issues (philosophical, theoretical, sociopolitical); (b) issues regarding methods and methodology; and (c)

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

contemporary applications of MMR. While not every chapter has a direct connection to education or assessment, there is much to be learned from the applications of MMR across these fields. The authors address emergent issues and present examples that provide learning opportunities about the different stances and purposes of MMR.

REFERENCES

Abbuhl, R., & Mackey, A. (2008). Second language acquisition research methods. In K. A. King & N. H. Hornberger (Eds.), *Encyclopedia of language and education: Research methods in language and education* (Vol. 10, pp. 1–13). Dordrecht, The Netherlands: Springer.

Anthony, J. J. (2009). Classroom computer experiences that stick: Two lenses on reflective timed essays. *Assessing Writing*, *14*, 194–205.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, *2*, 1–34.

Baker, B. A. (2010). Playing with the stakes: A consideration of an aspect of the social context of a gatekeeping writing assessment. *Assessing Writing*, *15*, 133–153.

Baker, B. A. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly*, *9*, 225–248.

Barkaoui, K. (2010). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly*, *44*, 31–57.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy, & Practice*, *18*, 279–293.

Bryman, A. (2007). Barriers to integrating quantitative and qualitative research. *Journal of Mixed Methods Research*, *1*, 8–22.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi- experimental designs for research*. Skokie, IL: Rand McNally.

Caracelli, V. J., & Greene, J. C. (1993). Data analysis strategies for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, *15*, 195–207.

Colby-Kelly, C., & Turner, C. E. (2007). AFL research in the L2 classroom and evidence of usefulness: Taking formative assessment to the next level. *The Canadian Modern Language Review/La Revue Canadienne des Langues Vivantes*, *64*, 9–37.

Cook, T. D. (1985). Postpositivist critical multiplism. In R. L. Shortland & M. M. Mark (Eds.), *Social science and social policy* (pp. 129–46). Newbury Park, CA: Sage.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Skokie, IL: Rand McNally.

Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281–302.

Datta, L. (1997). Multimethod evaluations: Using case studies together with other methods. In E. Chelimsky & W. R. Shadish (Eds.), *Evaluation for the 21st century: A handbook* (pp. 344–359). Thousand Oaks, CA: Sage.

Davies, A. (Ed.). (1997). Ethics in language testing [special issue]. *Language Testing*, *14*.

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

Dellinger, A. B., & Leech, N. L. (2007). Toward a unified validation framework in mixed methods research. *Journal of Mixed Methods Research*, *1*, 309–332.
Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods*. New York, NY: Praeger.
Derwing, T. M., Munro, M. J., & Thomson, R. I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics*, *29*, 359–380.
DiPardo, A., Storms, B. A., & Selland, M. (2011). Seeing voices: Assessing writerly stance in the NWP analytic writing continuum. *Assessing Writing*, *16*, 170–188.
Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. New York, NY: Oxford University Press.
Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, *13*, 208–238.
Gage, N. (1989). The paradigm wars and their aftermath: A "historical" sketch of research on teaching since 1989. *Educational Researcher*, *18*, 4–10.
Ghanbari, B., Barati, H., & Moinzadeh, A. (2012). Problematizing rating scales in EFL academic writing assessment: Voices from Iranian context. *English Language Teaching*, *5*, 76–90.
Greene, J. C. (2007). *Mixed methods in social inquiry*. San Francisco, CA: Jossey-Bass.
Greene, J. C. (2008). Is mixed methods social inquiry a distinctive methodology? *Journal of Mixed Methods Research*, *2*, 7–22.
Greene, J. C. (2011). The construct(ion) of validity as argument. In H. T. Chen, S. I. Donaldson, & M. M. Mark (Eds.), *Advancing validity in outcome evaluation: Theory and practice, New Directions for Evaluation* (pp. 81–92). San Francisco, CA: Jossey-Bass.
Greene, J. C., & Caracelli, V. J. (1997). Defining and describing the paradigm issues in mixed-method evaluation. In J. C. Greene & V. J. Caracelli (Eds.), *Advances in mixed-method evaluation: The challenges and benefits of integrating diverse paradigms* (pp. 5–18). San Francisco, CA: Jossey-Bass.
Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis*, *11*, 255–274.
Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, *6*, 427–439.
Guion, R. M. (1980). On trinitarian doctrines of validity. *Professional Psychology*, *11*, 385–398.
Habermas, J. (1984). *The theory of communicative action*. Boston, MA: Beacon Press.
Hamid, M., Sussex, R., & Khan, A. (2009). Private tutoring in English for secondary school students in Bangladesh. *TESOL Quarterly*, *43*, 281–308.
Harsch, C., & Martin, G. (2012). Adapting CEF-descriptors for rating purposes: Validation by a combined rater training and scale revision approach. *Assessing Writing*, *17*, 228–250.
Hashemi, M. R. (2012). Reflections on mixing methods in applied linguistics research. *Applied Linguistics*, *33*, 206–212.
Hesse-Biber, S., & Leavy, P (2006). Analysis and interpretation of qualitative data. In S. Hesse-Biber & P. Leavy (Eds.), *The practice of qualitative research* (pp. 343–374). London, UK: Sage.
Hyland, T. A. (2009). Drawing a line in the sand: Identifying the borderzone between self and other in EL1 and EL2 citation practices. *Assessing Writing*, *14*, 62–74.
Isaacs, T. (2008). Toward defining a valid assessment criterion of pronunciation proficiency in non-native English-speaking graduate students. *The Canadian Modern Language Review/La Revue Canadienne des Langues Vivantes*, *64*, 555–580.
Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility. *Studies in Second Language Acquisition*, *34*, 475–505.
Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for applying Fusion Model to LanguEdge assessment. *Language Testing*, *26*, 31–73.

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

Jang, E. E. (2013). Mixed methods research in SLA. In P. Robinson (Ed.), *The Routledge encyclopedia of SLA* (pp. 429–431). New York, NY: Routledge.
Jang, E. E., & Roussos, L. (2009). Integrative analytic approach to detecting and interpreting L2 vocabulary DIF. *International Journal of Testing*, *9*, 238–259.
Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, *33*, 14–26.
Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, *1*, 112–133.
Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527–535.
Kane, M. (2002). Validating high stakes testing programs. *Educational Measurement: Issues and Practice*, *21*, 31–41.
Kane, M. (2006). Validation. In R. Brennon (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.
Kim, Y. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, *26*, 187–217.
Kim, Y., & Jang, E. E. (2009). Differential functioning of reading subskills on the OSSLT for L1 and ELL students: A multidimensionality model-based DBF/DIF approach. *Language Learning*, *59*, 825–865.
Kirkhart, K. E. (2005). Through a cultural lens: Reflections on validity and theory in evaluation. In S. Hood, R. K. Hopson, & H. T. Frierson (Eds.) *The role of culture and cultural context: A mandate for inclusion, the discovery of truth, and understanding in evaluative theory and practice* (pp. 21–39). Greenwich, CT: Information Age.
Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*, 275–304.
Lee, Y., & Greene, J. (2007). The predictive validity of an ESL placement test. *Journal of Mixed Methods Research*, *1*, 366–389.
Lee, H., & Winke, P. (2013). The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test. *Language Testing*, *30*, 99–123.
Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.
Magnan, S. S. (2006). From the editor: The MLJ turns 90 in a digital age. *The Modern Language Journal*, *90*, 1–5.
Mathison, S. (1988). Why triangulate? *Educational Researcher*, *17*, 13–17.
Maxwell, J. A. (1992). Understanding and validity in qualitative research. *Harvard Educational Review*, *62*, 279–300.
Maxwell, J. A., & Loomis, D. M. (2003). Mixed methods design: An alternative approach. In A. Tashakkori & C. Teddlie (Eds.), *Sage handbook of mixed methods in social & behavioral research* (pp. 241–271). Thousand Oaks, CA: Sage.
Mertens, D. M. (2007). Transformative paradigm. *Journal of Mixed Methods Research*, *1*, 212–225.
Mertens, D. M. (2010). Philosophy in mixed methods teaching: The transformative paradigm as illustration. *International Journal of Multiple Research Approaches*, *4*, 9–18.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.
Miles, M. B., & Huberman, A. M. (1984). *Qualitative data analysis: A sourcebook of new methods*. Thousand Oaks, CA: Sage.
Mislevy, R. (1995). Test theory and language-learning assessment. *Language Testing*, *12*, 341–369.
Morgan, D. L. (2007). Paradigms lost and pragmatism regained. *Journal of Mixed Methods Research*, *1*, 48–76.
Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, *23*, 5–12.
Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing*, *28*, 483–508.

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

Nastasi, B. K., Hitchcock, J. H., & Brown, L. M. (2010). An inclusive framework for conceptualizing mixed methods design typologies: Moving toward fully integrated synergistic research models. In A. Tashakkori & C. Teddlie (Eds.), *Sage handbook of mixed methods in social and behavioral research* (pp. 305–338). Thousand Oaks, CA: Sage.

Onwuegbuzie, A. J., & Johnson, R. B. (2006). The validity issue in mixed research. *Research in the Schools*, *13*, 48–63.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.

Perrone, M. (2011). The effect of classroom-based assessment and language processing on the second language acquisition of EFL students. *Journal of Adult Education*, *40*, 20–33.

Plakans, L., & Gebril, A. (2012). A close investigation into source use in integrated second language writing tasks. *Assessing Writing*, *17*, 18–34.

Poonpon, K. (2011, April). "Synergy of mixed method approach to development of ESL speaking rating scale." Paper presented at Doing Research in Applied Linguistics [conference], Bangkok, Thailand.

Reichardt, D. S., & Cook, T. D. (1979). Beyond qualitative versus quantitative methods. In T. D. Cook & C. S. Reichardt (Eds.), *Qualitative and quantitative methods in evaluation research* (pp. 7–32). Thousand Oaks, CA: Sage.

Reiss, A. J. (1968). Stuff and nonsense about social surveys and participant observation. In H. S. Becker, B. Geer, D. Riesman, & R. S. Weiss (Eds.), *Institutions and the person: Papers in memory of Everett C. Hughes*. Chicago, IL: Aldine.

Schegloff, E. A. (1993). Reflections on quantification in the study of conversation. *Research on Language and Social Interaction*, *26*, 99–128.

Schwandt, T. A., & Jang, E. E. (2004). Linking validity and ethics in language testing: Insights from the hermeneutic turn in social science. *Studies in Educational Evaluation*, *30*, 265–280.

Shepard, L. A. (1992). What policy makers who mandate tests should know about the new psychology of intellectual ability and learning. In. B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessment: Alternative views of aptitude, achievement and instruction* (pp. 301–328). Boston, MA: Kluwer.

Tashakkori, A., & Teddlie, C. (Eds.). (2003). *Handbook of mixed methods in social and behavioral research*. Thousand Oaks, CA: Sage.

Tashakkori, A., & Teddlie, C. (Eds.). (2010). *Handbook of mixed methods in social and behavioral research* (2nd ed.). Thousand Oaks, CA: Sage.

Teddlie, C., & Tashakkori, A. (2003). Major issues and controversies in the use of mixed methods in the social and behavioral sciences. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 3–51). Thousand Oaks, CA: Sage.

Teddlie, C., & Tashakkori, A. (2006). A general typology of research designs featuring mixed methods. *Research in the Schools*, *13*, 12–28.

Turner, C. E. (2009). Examining washback in second language education contexts: A high stakes provincial exam and the teacher factor in classroom practice in Quebec secondary schools. *International Journal on Pedagogies and Learning*, *5*, 103–123.

Uchikoshi, Y., & Maniates, H. (2010). How does bilingual instruction enhance English achievement? A mixed-methods study of Cantonese-speaking and Spanish-speaking bilingual classrooms. *Bilingual Research Journal*, *33*, 364–385.

Weir, C. J. (2005). *Language testing and validation*. Basingstoke, UK: Palgrave Macmillan.

Winke, P. (2011). Evaluating the validity of a high-stakes ESL test: Why teachers' perceptions matter. *TESOL Quarterly*, *45*, 628–660.

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, *17*, 150–173.

Xi, X. (2010). Aspects of performance on line graph description tasks: Influenced by graph familiarity and different task features. *Language Testing*, *27*, 73–100.

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

Yin, M., Sims, J., & Cothran, D. (2012). Scratching where they itch: Evaluation of feedback on a diagnostic English grammar test for Taiwanese university students. *Language Assessment Quarterly*, *9*, 78–104.
Yu, G. (2010). Effects of presentation mode and computer familiarity on summarization of extended texts. *Language Assessment Quarterly*, *7*, 119–136.

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

APPENDIX A

Summary of Criteria Used for Identifying MMR Studies

Research Category	Details
Inclusion (exclusion) criteria	Peer-reviewed research in the field of language assessment;
	Written in English;
	Focus is assessment of English as a second/foreign language (assessment of English language proficiency);
	Population of interest is all ages;
Key terms	Mixed methods
	Mixed-method
	Mixed-methods
	Mixed methodology
	Multiple methods
	Triangulation
	Language assessment
	Language learning
	Assessment and evaluation; language test/s/ing
	Integration
Databases	ERIC (ProQuest)
	Education Full Text
	CBCA Education
	ProQuest Education Journals
	Social Sciences Citation Index
	FRANCIS (Humanities & Social Sciences)
	Linguistics & Language Behavior Abstracts (LLBA)
Journals	*Language Testing*
	Language Assessment Quarterly
	Assessing Writing
	Assessment in Education: Principles, Policy & Practice
	Assessment and Evaluation in Higher Education
	Educational Evaluation and Policy Analysis
	Educational Measurement: Issues and Practice
	Educational and Psychological Measurement
	Evaluation and Research in Education
	Journal of Educational Measurement
	Practical Assessment, Research and Evaluation
	Applied Psychological Measurement
	Psychometrika
	TESOL Quarterly
	Language Learning
	Modern Language Journal
	Canadian Modern Language Review
	Studies in Second Language Acquisition

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press

APPENDIX B

Description of Categories of Analysis

Category	Description
Population of interest	The population about whom we are concerned and about whom we would like to draw inferences, or who are impacted by the research. They are not necessarily the focus of the data collection, but the research will inform this population.
Study participants	The people who participated in the study and from whom data was collected
Language modality	The specific language skills on which the article focused (e.g., listening, oral communication skills, writing)
Research context	The language learning and assessment situation in which the study is based (e.g., classroom-based language testing, EFL)
Research questions and/or purpose(s)	The questions and/or purposes as defined by the researcher(s)
Timing of mixing	Refers to the stage in the research when mixing happened (e.g., developmental stage, analysis, discussion)
Rationale for use of mixed methods	An explanation of the reasons why mixed methodology was employed. Reasons (explicit or implicit) were categorized according to three purposes: complementarity, triangulation, transformative
Data sources	The types of data gathered to respond to the research questions and/or purpose(s)
Data analysis approaches	Types of analyses conducted (e.g., statistical analyses, quantization of qualitative data)
Presentation of results	The organization of the results and the data sources and analyses used to present them
Synergetic effects	Examination of the research to determine if the effect of using mixed method results in a research process and findings that are greater than if mixed method had not been employed (Nastasi et al., 2010)

https://doi.org/10.1017/S0267190514000063 Published online by Cambridge University Press