

# **Influential choices: deconstructing operationalisations of indigeneity in survey-based education research using an example from Peru**

**Authors:** Miriam Broeks and Ricardo Sabates Aysa

## **About the authors:**

Miriam Broeks is a PhD candidate at the Faculty of Education, University of Cambridge, where she is a member of the Research for Equitable Access and Learning (REAL) Centre. Her research focuses on exploring the educational realities of indigenous pupils in Peru. She is also Senior Analyst at RAND Europe. ([mglb3@cam.ac.uk](mailto:mglb3@cam.ac.uk))

Dr Ricardo Sabates Aysa is Professor of Education and International Development at the University of Cambridge, and a member of the Research for Equitable Access and Learning (REAL) Centre. His research focuses on educational inequalities in access and learning in some countries of the Global South using a lifecourse approach. ([rs867@cam.ac.uk](mailto:rs867@cam.ac.uk))

**Abstract:** Indigeneity is a complex social construct that can be defined in multiple ways using diverse markers traditionally based on the characteristics of individuals. Survey-based studies have used language, self-identification or location information to operationalise indigeneity. Yet, as suggested by Walters and Andersen (2013), Gillborn et al. (2018) and others, few scholars reflect on how the indigeneity variable is specified and whether this operationalisation may impact results. This paper examines this issue empirically using the case of indigeneity in Peru. First, survey-based empirical studies are identified to explore the ways in which indigeneity has been operationalised. Then, using the Young Lives study we present diverse operationalisations of indigeneity and outline how these may lead to different educational outcomes for children. We show that quantitative researchers using survey-based data should engage more deeply with different operationalisation of indigeneity as these can lead to different educational outcomes for children categorised as Indigenous.

**Key words:** indigeneity, Peru, quantitative methods, education, Young Lives survey

## **Introduction**

Various empirical studies using survey-based data show that Indigenous pupils in Peru experience notable educational inequalities compared to their non-Indigenous counterparts (Ames, 2012; Arteaga & Glewwe, 2014; Hernandez-Zavala et al., 2006; Hynsjö & Damon, 2016; Skoufias et al., 2010; Van der Tuin & Verger, 2013). While many factors have been identified as drivers or mediators that explain these inequalities, it is rare to find studies that consider whether the operationalisation of indigeneity itself may impact on the results obtained. For instance, it may be that Indigenous pupils as defined by language may have different educational outcomes than Indigenous pupils as defined by ethnic origin. The operationalisation of indigeneity matters not just for how Indigenous groups are defined and by whom they are defined, but results emerging from survey-based data may be used to inform educational policies. In particular, it is argued that survey-based data which is not critically examined is likely to reinforce the position of exclusion that Indigenous groups have historically experienced (Treviño Villarreal, 2006; Walter, 2010).

This paper aims to show the complexity around defining indigeneity as a variable in quantitative surveys and highlights the importance of reflecting on how the concept is operationalised through empirical research. It takes principles from Indigenous Statistics literature and Quantitative Critical Race Theory (QuantCrit) as inspiration to review previous quantitative empirical studies and uses data from the Young Lives study in Peru to provide a concrete example. First, we explore whether empirical studies offer different options to operationalise indigeneity or whether they discuss the limitations around how the data about Indigenous people was collected in quantitative studies in Peru. To clarify our position, ‘operationalising indigeneity’ refers to computing an indigeneity variable using information from an existing survey-based dataset. Second, we aim to show that quantitative researchers should engage more deeply with different operationalisation of indigeneity as these can lead to different educational outcomes for children categorised as Indigenous. A deeper engagement can shed light on processes that may explain why differences are observed.

## **Relevant theoretical underpinnings to understand indigeneity issues**

Indigeneity is a complex concept (Snipp, 2016). Literature on the subject showcases a myriad of possible aspects that can constitute Indigenous identity (Greene, 2016; Levi & Maybury-Lewis, 2012; Paredes, 2007; Vásquez Huamán et al., 2012; Walter & Andersen, 2013; Williams & Schertzer, 2019). The key markers, as synthesised by the United Nations (2004), include historical antecedence, occupation of ancestral land, ethnic identity, cultural distinctiveness (dress, lifestyle, community membership), language, self-identification, and non-dominance in relation to a colonial nation state.

The diversity of markers demonstrates the multifaceted nature of indigeneity, yet not all these markers are available in a single survey-based study. Scholars such as Gillborn et al. (2018), Tuhiwai Smith

(2013), Walter (2010), Walter and Andersen (2013), and Walter and Suina (2019) have raised the question as to which markers should be included in surveys when exploring Indigenous issues and emphasise the need to critically reflect on why certain markers are included in surveys. In the case of Peru, and as identified through the literature review conducted as part of this study, many survey-based studies have introduced markers of indigeneity based on language or self-identification to establish differences between Indigenous and non-Indigenous groups. However, some point to disadvantages of using these markers to establish indigeneity in Peru given the complex historical, cultural, political and societal factors unique to the Peruvian context. For example, discrimination against Indigenous populations in Peru is rooted in colonial power imbalances and can lead to language loss and avoidance of self-identifying as Indigenous to evade stigma (Cánepa, 2008; Figueroa & Barrón, 2005; Paredes, 2007; Zarate, 2011). The multifaceted nature of indigeneity also raises the question as to whether researchers should use a single indicator or multiple indicators when establishing the Indigenous status of an individual for analyses. Unfortunately, as already discussed by Walters and Andersen (2013) and Gillborn et al. (2018), many studies using survey-based data that explore matters related to Indigenous people do not provide a detailed account of how they identified individuals as Indigenous.

In order to overcome some of these limitations and to engage more critically with the ways in which data is collected, and social concepts operationalised and analysed, Gillborn et al. (2018) put forward five principles to guide quantitative research; two of which inspire this paper. First, data is often collected and analysed in ways that replicate the assumptions of the dominant group (Walter, 2010). Therefore, results should be examined thinking about what role they may play in promoting “deficit analyses” (Gillborn et al., 2018, p. 158). Second, social categories by themselves may not be the sole cause of disadvantage, but it is the process that leads to these disadvantages which must be examined as the root causes of unequal outcomes (Gillborn et al., 2018; Walter & Suina, 2019). In this paper we reflect on how different survey-based studies using Peruvian data have engaged with the concept of indigeneity and we investigate whether diverse operationalisations of indigeneity matter when measuring educational outcomes. It is important to highlight that we do not delve into in-depth analyses of how educational outcomes in the Young Lives dataset were operationalised, nor with understanding the process or mechanisms that can explain differences. However, we acknowledge that while we may consistently find that Indigenous children obtain lower scores in educational attainment tests (we find ‘deficit’), it is important to establish a narrative beyond deficit in line with QuantCrit and Indigenous Statistics principles.

## **Methodology**

We write about Indigenous populations in Peru given our Latin American roots and interest in promoting equity through education and in breaking colonial power imbalances that are reinforced through research. We are both mestizos although we would be classified as ‘White, other’ in Western

contexts. One of us is Peruvian-Dutch, born in the Andean highlands of Peru. The other is Mexican, with an ethnicity that combines Spanish, Lebanese and mestizo background. While we acknowledge the place of privilege we write from, both working in UK academia, we are interested in making a contribution to encourage reflection in educational research.

Two overarching questions guide this research: 1) how have others operationalised indigeneity in Peru? And 2) does the chosen operationalisation of indigeneity influence observed differences in educational attainment between Indigenous and non-Indigenous children? Empirically, two steps are taken to establish and examine different operationalisations of indigeneity: first, we conduct a targeted review of published empirical survey-based studies reporting on quantitative analyses using Young Lives Peru and other Peru survey-based data. Second, we conduct statistical analyses of different operationalisations of indigeneity linked to educational outcomes using Young Lives Peru data.

The targeted review of the survey-based literature was conducted to identify how quantitative researchers using data from Peru operationalise pupils' Indigenous status. Systematic searches were conducted using Boolean terms on Scopus, ERIC, Web of Science and Google Scholar. All searches were conducted in English and Spanish using translations of the search terms.<sup>1</sup> Academic sources (journal articles) and grey literature (i.e. government reports, websites of Indigenous groups, dissertations, reports from international organisations and research institutes, etc.) were included. We actively searched for research conducted by Peruvian Indigenous scholars. However, we did not identify studies conducted by Peruvian researchers that self-identify as Indigenous. Therefore, the research presented here has been carried out by Peruvian scholars that research Indigenous issues.

The literature review yielded one-hundred documents that met the inclusion criteria and that were examined, fifty-one were studies using Young Lives data, forty-nine used other social survey data from Peru. Information was systematically recorded using an extraction template and thematically analysed to establish whether and, if so, how Indigenous pupils were defined in each study. This approach allowed to identify thematic patterns across the different studies. Findings from the review informed how to compute the indigeneity variables for this study.

Following the review of studies, we use the Young Lives Peru study to operationalise indigeneity in multiple ways and to examine if such operationalisations matter for investigating educational outcomes of Indigenous children. The Young Lives study is a longitudinal survey that has collected five rounds of data. Young Lives is among the most complete sources of social research data which have been collected in Peru and one which contains several indicators that relate to the sociodemographic status of the households where selected children live (Young Lives, 2017a). Children were selected for the study using a multi-stage, cluster-stratified, random sampling approach, oversampling for poor areas (Young Lives, 2017b). To engage with the data, we utilise information only from the young cohort as this has more indicators related to the operationalisation of indigeneity.

For academic attainment, we use results from one of the education outcomes available when these children were around 12 years of age, namely the Peabody Picture Vocabulary Test (PPVT).

It is important to explain how the PPVT was collected, and whether the ways in which the test is designed may carry bias against Indigenous pupils. While we are unable to control for any of these potential biases, it is our duty to reflect about these when describing measures in social research. The PPVT is a standardised vocabulary test to assess verbal ability and receptive vocabulary of children. It lasted between 20 to 30 minutes to implement. Children were shown groups of four pictures. Then, the test administrators said a word and the child was asked to identify the picture that best corresponded to the word. In Peru, the Young Lives study used the Spanish translated version of PPVT-R (a revised version of the first published PPVT). Children were invited to respond to the test in the language they felt most comfortable (e.g. Quechua or, Spanish and Quechua) even if the Spanish version of PPVT-R was used (Cueto & León, 2012). Following Stein and Lukasik (2009), an advantage of using PPVT is that it does not require the child to be able to read, write or talk. This informed our decision to focus on PPVT and not use the mathematics or reading comprehension test data also available for round 4; a limitation of these outcome measures is that they require that the child is able to read.

It is worth noting that Peru is an ethnically and linguistically diverse country. According to the 2017 national census, six million people self-identified with an indigenous ethnicity, and four million respondents reported speaking an Indigenous language (Hospina, 2019). While most of the Indigenous population speaks Quechua (83 percent), other Indigenous languages include Aymara (11 percent) and native languages from the amazon (6 percent) (Guerrero et al., 2012; International Work Group on Indigenous Affairs, 2021).

Walter and Andersen (2013) reiterate the importance of reflecting on potential limitations of outcome measures used in research. While PPVT can be considered the best available test for our purposes, we are aware that it has some limitations. Cueto and León (2012) and Dawes (2020) note that the PPVT test was not standard across languages or countries and advise caution when comparing PPVT scores of pupils speaking different languages.

In addition, after consulting the Young Lives team, it was established that the Spanish PPVT-R was not officially translated for administration in Indigenous languages, although fieldworkers may have spoken an Indigenous language when administering the test if requested by the child. The lack of official translation into Indigenous languages is a crucial issue that has implications for the validity of the test. Further, it remains unclear whether Indigenous individuals were involved in the process to develop guidelines for administering PPVT with Indigenous children whose first language is not Spanish. Involving Indigenous representatives is a key aspect that proponents of Indigenous methods highlight as important to ensure the integrity and respect of Indigenous knowledges when conducting

research about them (Hutchinson et al., 2014; Tuhiwai Smith, 2013).<sup>2</sup> Apart from this, to the best of our knowledge the PPVT test was not adapted to include words and images which would be equally familiar to all children regardless of cultural background. For instance, it could be that the image of a maize plant is more familiar for rural children than urban children, whereas an image of a building may be more familiar for urban than rural children.

In terms of the data collection process, Young Lives recruited fieldworkers with different language skills. In the case of Peru, this meant fieldworkers speaking Spanish and Quechua (Oré et al., 2012). However, there are children in the Young Lives Peru dataset whose language is other than Spanish or Quechua. It remains unclear whether fieldworkers matching these language skills were also recruited. All fieldworkers received training and participated in the piloting of the survey in urban and rural areas, which involved practice sessions and discussions around the correct translation of complex questions into local languages (Oré et al., 2012; Young Lives, 2017c). Manuals and protocols were available to facilitate the correct administration of the instruments. More information about the procedures followed by Young Lives can be found online.<sup>3</sup>

The analysis of the different indigeneity operationalisations is based on boxplots mapping each indigeneity variable against the PPVT scores. To obtain the statistical significance of differences of educational outcomes between different operationalisations of indigeneity we use two sample *t*-tests. All estimations use the non-Indigenous group as comparison. Standard errors, the 95% Confidence Interval and *p*-value are provided.

To operationalise indigeneity, we started by using some of the key variables available in the Young Lives survey. The child's and mother's first language and ethnicity (available from round 2 of Young Lives) were recoded to classify children into Indigenous and non-Indigenous groups.

Table 1 shows how this information was recategorised. Children speaking Quechua, Aymara and a language from the jungle are grouped together in the Indigenous category. In addition, it was decided to give precedence to speaking an Indigenous language and therefore, children that speak both Spanish and Quechua, or Spanish and Aymara, were also included within the Indigenous category. In terms of the ethnicity variable, children categorised as Asian or Afro-descendants are excluded from analyses as they represented a small number of children (i.e. 22 or 1.2% of cases from round 2) (Cueto et al., 2019; Figueroa, 2003; Heckert, 2010).

**Table 1. Recoding of mother and child first language and ethnicity information**

<b>Original variable categories</b>	<b>Recoded variable categories</b>
First language (of child or mother) <ul style="list-style-type: none"> <li>- Spanish,</li> <li>- Quechua,</li> <li>- Aymara,</li> <li>- Native language from jungle,</li> <li>- Nomatsiguenga (native language in jungle),</li> <li>- Spanish and Aymara,</li> <li>- Spanish and Quechua,</li> <li>- Mute or difficulty speaking</li> <li>- "nk"</li> <li>- "n/a"</li> <li>- "refused to answer",</li> <li>- "missing"</li> </ul>	First language (of child or mother) <ul style="list-style-type: none"> <li>- Indigenous (includes:               <ul style="list-style-type: none"> <li>o Quechua</li> <li>o Aymara</li> <li>o Native language from jungle,</li> <li>o Nomatsiguenga (native language in jungle),</li> <li>o Spanish and Aymara,</li> <li>o Spanish and Quechua)</li> </ul> </li> <li>- Non-Indigenous (includes: Spanish)</li> </ul> <i>Removed from analyses:</i> <ul style="list-style-type: none"> <li>- Mute or difficulty speaking</li> <li>- "nk", "n/a"</li> <li>- "refused to answer"</li> <li>- "missing"</li> </ul>
Ethnicity (of child or mother) <ul style="list-style-type: none"> <li>- White,</li> <li>- Mestizo,</li> <li>- Quechua,</li> <li>- Aymara,</li> <li>- Amazon indian,</li> <li>- Negro, mulato, zambo,</li> <li>- Asian/oriental",</li> <li>- "nk",</li> <li>- "n/a",</li> <li>- "refused to answer",</li> <li>- "missing"</li> </ul>	Ethnicity (of child or mother) <ul style="list-style-type: none"> <li>- Indigenous (includes: Quechua, Aymara, Amazon Indian)</li> <li>- Non-Indigenous (includes: white, mestizo)</li> </ul> <i>Removed from analyses:</i> <ul style="list-style-type: none"> <li>- Negro, mulato, zambo,</li> <li>- Asian/oriental</li> <li>- "nk", "n/a"</li> <li>- "refused to answer"</li> <li>- "missing"</li> </ul>

*Source: Authors' creation*

Informed by the targeted literature review, ten different indigeneity variables are computed. Four are based on binary categorisations of Indigenous and non-Indigenous described in



Table 1. Six of these variables are multi-category and include more granular groupings. More details on how these variables were computed and the rationale behind our choices is presented after the review of studies.

## **Results and discussion**

### **Operationalisation of indigeneity in past research**

#### *Using language information to operationalise indigeneity*

The literature review revealed that most screened studies used language to identify Indigenous individuals. Some authors using language to operationalise indigeneity in their studies explained the rationale for their choice. Some claimed that it is the most ‘objective’ approach compared to, for example self-identification, on the basis that discrimination and prejudice can lead individuals to be less likely to self-identify as Indigenous (Arteaga & Glewwe, 2014; Cueto, Miranda, et al., 2016). Others justified their choice based on what is most commonly used in Peru from a governmental perspective, drawing on the fact that the Peruvian government has mainly used language to measure the size of the Indigenous population in national statistics (Pasquier-Doumer & Risso Brandon, 2015). This has indeed been a trend in official Peruvian governmental statistics up until recently (Alcázar, 2019; Arteaga & Glewwe, 2014; Telles, 2014). However, this trend is starting to change - since 2017 questions on self-identification are being integrated into national surveys and the national census (Instituto Nacional de Estadística e Informática, 2018).

While most reviewed studies using Young Lives data based indigeneity categorisations on language, there is considerable variation in what information about language is used. From the 51 articles using Young Lives Peru data, 34 referred to language aspects. Within these 34 articles multiple aspects of language were used, namely: the ‘child’s mother tongue’, the ‘mother’s mother tongue’, the first language of the father, caregiver, or of the ‘parents or grandparents’. The most common approach was using ‘mother’s first language’, either in isolation or in combination with other elements such as the child’s language, language of another parent and or self-identification.<sup>4</sup>

Studies by Cueto, Escobal et al. (2011), Cueto, Guerrero et al. (2014), Cueto, Singh et al. (2016) and Pasquier-Doumer and Risso Brandon (2015) exemplify the variety of approaches used to categorise children’s indigeneity. Cueto and colleagues used child and parent or family level language information in their studies. Indigenous and non-Indigenous children were identified based on ‘mother tongue at home’ specified as “the first language of the child is Indigenous or both parents speak an Indigenous language but the child does not” (Cueto, Guerrero et al., 2014, p. 251); according to ‘Indigenous home’ using “mother’s first language” (Cueto, Escobal et al., 2011, p. 42); or by ‘minority language at home’ when “at least one parent and the child had an Indigenous mother tongue” (Cueto, Singh et al., 2016, p. 7). The study by Pasquier-Doumer and Risso Brandon (2015)

used two definitions of Indigenous children: one based on ‘the language of a parent (either mother, father or caregiver)’, and another on ‘the language in which the mother spoke to her child’.

The review of studies using other social survey data from Peru revealed a similar trend. Many (24 articles of 49) used language to define indigeneity. In this case too, different aspects of language were used (main language in the household, whether student grew up speaking an Indigenous language, whether the child’s mother speaks an Indigenous language, etc.) (Flores-Bendezú et al., 2015; Ibarra, 2019; Sakellariou, 2008; UNICEF, 2010; Valdivia, 2007; Vásquez Huamán et al., 2012; Yamada et al., 2011).

Valdivia (2007) specified three variables to establish indigeneity using data from the Family Health and Demography Survey (ENDES) labelled as follows:

- Quechua/Aymara speaker – women who identify these languages as the main one they use at home.
- Recent Spanish speaker – women who report using Spanish most often at home but identify Quechua or Aymara as the mother tongue of their parents or grandparents.
- Old Spanish speaker – women who report using Spanish at home and that it is the mother tongue of their parents and grandparents.

These variables yielded different proportions of Indigenous individuals in the sample, with the ‘Quechua/Aymara speaker’ definition giving the smallest grouping.

### *Combining language and self-identification information*

Acknowledging the complexity to operationalise indigeneity using quantitative data, other studies combined information on language with self-identification. The past two decades have seen a transformation in the conceptualisation of ethnicity in Peru, one which highlights the importance of self-identification to assert the rights and recognition of Indigenous populations (Alcázar, 2019; Delprato, 2019; Instituto Nacional de Estadística e Informática, 2018; Telles, 2014; Valdivia Vargas, 2011; Zarate, 2011). This is reflected in changes in governmental statistics described above and in various of the sources reviewed.

Three of the reviewed studies that used language and self-identification information to define indigeneity provided details on their approach. Castro and Yamada (2010) gave precedence to ‘mother tongue’ over ‘self-identification’ to define Indigenous individuals in their sample (using National Household Survey (ENAHU) data). They classified as ‘Indigenous’ individuals whose mother tongue was Quechua or Aymara irrespective of their self-identification. However, for those who reported having Spanish as their mother tongue, their self-identification was used to determine their ethnicity.

Delprato (2019) and Trivelli (2005) also used language and self-identification dimensions. Using data from the Third Regional Comparative and Explanatory Study (TERCE), Delprato (2019) specified five possible definitions for ‘Indigenous’ students as follows: 1) both parents are Indigenous, 2) one parent is Indigenous, 3) the language parents speak to the child, 4) mother is Indigenous, father is non-Indigenous, and 5) student self-identifies as Indigenous. Trivelli (2005) also specified five possible definitions for indigeneity using ENAHO data. Namely, Trivelli identified Indigenous individuals based on: 1) mother tongue is other than Spanish, 2) self-identification, 3) definition 1 and/or 2, 4) most frequently used language is other than Spanish, 5) parents or grandparents of the head of the household or spouse had a mother tongue other than Spanish.

### *Other markers used to operationalise indigeneity*

A small number of screened studies reported using geography and race information to define indigeneity. Figueroa and Barrón (2005), used place of origin to determine indigeneity, arguing that Peruvian history on rural-urban migration support this choice. They used seven categories to map three ethnic groups (white, mestizo and Indigenous) across the three geographic regions of Peru.

Paredes (2007) used data from the Centre for Research on Inequality, Human Security and Ethnicity (CRISE) study and took information on race (self-perceived skin colour), geography (place of origin), language and self-identification to develop four approximations of indigeneity. The study explored how indigeneity based on race, geography and language compared to self-identification answers from surveyed individuals. Paredes found that geography (following the same approach used by Figueroa and Barrón (2005)) and race gave a closer approximation to self-identification answers than language (i.e. how geography links to self-identification compared to how language does).

In the Project on Ethnicity and Race in Latin America (PERLA) indigeneity was conceptualised as multifaceted, conformed by personal and external influences, namely self-identification, ascription (attribution by others) and language. They measured ethnicity using six different indicators: 1) a classification ascribed by the interviewer using six categories from the Latin American Public Opinion Project (LAPOP) (i.e. Indigenous, *mestizo*, white, *mulato*, black, other), 2) self-identification using categories from LAPOP, 3) self-identification based on racial origins (open-ended question), 4) ethnic/racial self-identification using the format used in the ENAHO survey (on perceived belonging to a group based on customs and ancestors), 5) respondents mother tongue, and 6) a classification ascribed by the interviewer using a colour palate (Telles, 2014; Valdivia Vargas, 2011). These different definitions yielded markedly different proportions of Indigenous individuals in the sample (ranging from under 5% to nearly 40%).

## **Evidence on differing results depending on indigeneity operationalisation**

Consistent with concerns voiced by Indigenous Statistics and QuantCrit scholars, from the one-hundred studies reviewed, sixty-four provided a definition of how Indigenous cases were identified. Of these, twenty-eight explained the rationale behind their choice. Further, only sixteen of the one-hundred reviewed articles considered more than one marker (e.g. language, self-identification) to establish their definition of indigeneity, and only five of the studies developed different definitions of indigeneity, ran separate analyses by definition and reported on whether using different definitions influenced results (albeit providing varying levels of information).

In only one study the definition used did not seem to matter. Pasquier-Doumer and Risso Brandon (2015) studied student aspirations and educational investment (i.e., level of effort placed on educational attainment) and ran analyses with the two indigeneity definitions they developed.<sup>5</sup> Their results were consistent regardless of definition used.

In the case of Valdivia (2007), differences by indigeneity definition were found for some but not all the maternal health variables analysed. Statistically significant differences between definitions were found when looking at family planning and pregnancy patterns even after controlling for confounders. Those identified as Indigenous in terms of being ‘Quechua/Aymara speakers’ made the least use of services compared to the non-Indigenous group. Differences by definition were also found on other metrics, showing that disadvantages affect mainly those who predominantly speak Indigenous languages. However, these differences were conditioned out once other factors are included in the estimation (i.e. after controlling for education, socioeconomic status of the household and access to health infrastructure).

Two other studies also found differing results depending on indigeneity definition used. Delprato (2019) examined parental expectations, children’s maths and reading achievement, and found that the Indigenous group defined using the third definition based on language (i.e. the language parents speak to the child) was the most disadvantaged. These children had lower scores and their parents reported lower expectations. Trivelli (2005) ran analyses using several definitions and also found differences. Their analyses showed that within the Indigenous groups, those defined using language most frequently used were at further disadvantage in terms of socio-economic status compared to the other Indigenous groups.

Finally, Telles (2014) concluded that the choice of indigeneity variable influenced their results looking into how different definitions predicted ethno-racial inequality and experiences of discrimination. However, the study did not provide a full account of how.

## Examining different operationalisations of indigeneity using Young Lives data

Three key takeaways can be drawn from the above evidence that are helpful for shaping the analyses of the Young Lives dataset. First, from the screened studies most use a single factor to operationalise indigeneity, with language information being predominantly used. Yet, many of the available survey-based datasets, including Young Lives, contain more than one indicator that can be used to define it. Second, few of the identified studies outline their rationale behind the operationalisation of indigeneity they use. Justifying the operationalisation of indigeneity is important in light of the various possibilities survey-based datasets provide for this. Third, of the studies that conducted analyses using multiple indicators, most show that the chosen operationalisation matters (Walter, 2010). It stands out that children defined as Indigenous according to the language they speak face the greatest disadvantage, potentially due to the challenges they face in education and society and not due to their inability to achieve.

Based on these findings, we use Young Lives to provide as broad an operationalisation of indigeneity as possible. The Young Lives Peru dataset contains several indicators (e.g. language, ethnicity and location) that can be used in isolation or combination to categorise children according to indigeneity. In this paper we focus on language and ethnicity of the child and the mother<sup>6</sup> and developed binary and multi-category variables. The multi-category variables were informed by the reviewed literature that highlights the complexity of the concept of indigeneity. They attempt to address the multifaceted nature of indigeneity within the limits of the available data. It was decided to use both information on the child and the mother as these complement each other. More specifically, information from the mother is explored to tap on to aspects around ‘ancestry’ and informed by literature suggesting that speaking Indigenous languages can be lost over time by younger generations. Information on location is not used within these operationalisations as the Young Lives dataset does not include detailed enough information to identify municipalities as Indigenous or non-Indigenous. Only broader markers at the region and rural-urban level are available.

Table 2 provides an overview of the ten indigeneity variables that are explored. The first four are binary and are based on a single variable from the Young Lives dataset (*Indg1*, *Indg2*, *Indg3* and *Indg4*). These are computed given that the majority of reviewed articles use binary variables. A further six variables that combine language and or ethnicity information from the child and mother are also explored. These are henceforth referred to as multi-category variables. The fifth and sixth variables allocate cases into three groups. *Indg5* is computed by giving precedence to language (i.e. category one includes cases where the child and the mother speak an Indigenous language regardless of ethnicity, category two includes children that speak Spanish but their mother an Indigenous language, and category three cases where the child and the mother speak Spanish). The same rationale is followed for *Indg6* but using ethnicity (of the child and the mother) as the guiding marker.

Variables *Indg7* to *Indg10* combine language and ethnicity information and have four categories, two which are the same across them. These are the categories labelled as ‘Indigenous’ and ‘non-Indigenous’ meaning that cases are either Indigenous or non-Indigenous across all markers (i.e. for language and ethnicity of the child and mother). For *Indg7* and *Indg8* the language and ethnicity of the *child* led the categorisation respectively, while for *Indg9* and *Indg10* the language and ethnicity of the *mother* guided the groupings.

To see if the operationalisation of indigeneity matters, we run analyses using an educational attainment indicator based on PPVT test scores to distinguish between Indigenous and non-Indigenous children’s results. Table 3 provides information on the average percentage difference in PPVT scores between groups for the different indigeneity variables, using the non-Indigenous group as the reference group.

**Table 2. Overview of indigeneity variables examined**

Variable	Categories - Name (n)	Marker(s) based on
<b>Indg1</b>	1. Indigenous (252) 2. Non-Indigenous (1,551)	Child language
<b>Indg2</b>	1. Indigenous (537) 2. Non-Indigenous (1,266)	Mother language
<b>Indg3</b>	1. Indigenous (288) 2. Non-Indigenous (1,515)	Child ethnicity
<b>Indg4</b>	1. Indigenous (425) 2. Non-Indigenous (1,378)	Mother ethnicity
<b>Indg5</b>	1. Indigenous (child and mother speak an Indigenous language) (252) 2. Child language is non-Indigenous, but mother's language is Indigenous (287) 3. Non-Indigenous (child and mother speak Spanish) (1,264)	Combination between child and mother language and ethnicity. Language is given precedence over ethnicity to determine belonging to the middle group.
<b>Indg6</b>	1. Indigenous (child and mother are ethnically Indigenous) (288) 2. Child-ethnicity is non-Indigenous, but mother's ethnicity is Indigenous (143) 3. Non-Indigenous (child and mother are ethnically non-Indigenous) (1,372)	Combination between child and mother language and ethnicity. Ethnicity is given precedence over language to determine belonging to the middle group.
<b>Indg7</b>	1. Indigenous <i>across all markers</i> * (134) 2. Child language is Indigenous (regardless of other categories) (118) 3. Child language is non-Indigenous (regardless of other categories) (392) 4. Non-Indigenous <i>across all markers</i> (1,159)	Combination between child and mother language and ethnicity, with child language leading the categorisation.
<b>Indg8</b>	1. Indigenous <i>across all markers</i> * (134) 2. Child ethnicity is Indigenous (regardless of other categories) (154) 3. Child ethnicity is non-Indigenous (regardless of other categories) (356) 4. Non-Indigenous <i>across all markers</i> (1,159)	Combination between child and mother language and ethnicity, with child ethnicity leading the categorisation.
<b>Indg9</b>	1. Indigenous <i>across all markers</i> * (134) 2. Mother language is Indigenous (regardless of other categories) (291) 3. Mother language is non-Indigenous (regardless of other categories) (219) 4. Non-Indigenous <i>across all markers</i> (1,159)	Combination between child and mother language and ethnicity, with mother language leading the categorisation.
<b>Indg10</b>	1. Indigenous <i>across all markers</i> * (134) 2. Mother ethnicity is Indigenous (regardless of other categories) (291) 3. Mother ethnicity is non-Indigenous (regardless of other categories) (219) 4. Non-Indigenous <i>across all markers</i> (1,159)	Combination between child and mother language and ethnicity, with mother's ethnicity leading the categorisation.

\* Indigenous *across all markers* means that the child and mother speak an Indigenous language and have an Indigenous ethnicity. The opposite applies for non-Indigenous across all markers.

Source: Authors' creation

**Table 3. Independent sample t-test results for indigeneity variables (reference group ‘non-Indigenous’)**

Variable type	Variable name		Group	PPVT		
				Estimate	Std. error	p-value
Binary	Indg1	Indigenous		-15.12	0.85	13.45 - 16.79
	Indg2	Indigenous		-9.01	0.88	7.26 - 10.76
	Indg3	Indigenous		-9.70	0.71	8.30 - 11.09
	Indg4	Indigenous		-6.82	0.79	5.27 - 8.37
3-categories	Indg5	Indigenous		-15.91	0.86	14.21 - 17.60
		Child language is non-Indigenous, mother’s language is Indigenous		-4.26	0.87	2.55 - 5.97
	Indg6	Indigenous		-9.18	0.90	7.42 - 10.94
		Child-ethnicity is non-Indigenous, mother’s ethnicity is Indigenous		-1.82	1.24	-0.62 - 4.26
4-categories	Indg7	Indigenous		-16.70	1.03	14.67 - 18.73
		Child language is Indigenous		-14.83	1.32	12.22 - 17.44
		Child language is non-Indigenous		-2.78	0.78	1.26 - 4.30
	Indg8	Indigenous		-16.70	1.03	14.67 - 18.73
		Child ethnicity is Indigenous		-4.90	1.17	2.58 - 7.21
		Child ethnicity is non-Indigenous		-5.86	0.86	4.16 - 7.56
	Indg9	Indigenous		-16.70	1.03	14.67 - 18.73
		Mother language is Indigenous		-7.28	0.81	5.69 - 8.86
		Mother language is non-Indigenous		0.86	1.30	-3.44 - 1.73
	Indg10	Indigenous		-16.70	1.03	14.67 - 18.73
		Mother ethnicity is Indigenous		-4.04	0.90	2.26 - 5.81
		Mother ethnicity is non-Indigenous		-7.60	1.06	5.51 - 9.69

*Note: Values under ‘Estimate’ column are the percentage difference ( $\Delta$ ) between the group and the intercept (non-Indigenous group). These were computed using Welch Two Sample t-test. CI = Confidence Interval.*

*Source: Authors’ creation*



### *Examining binary variables against PPVT scores*

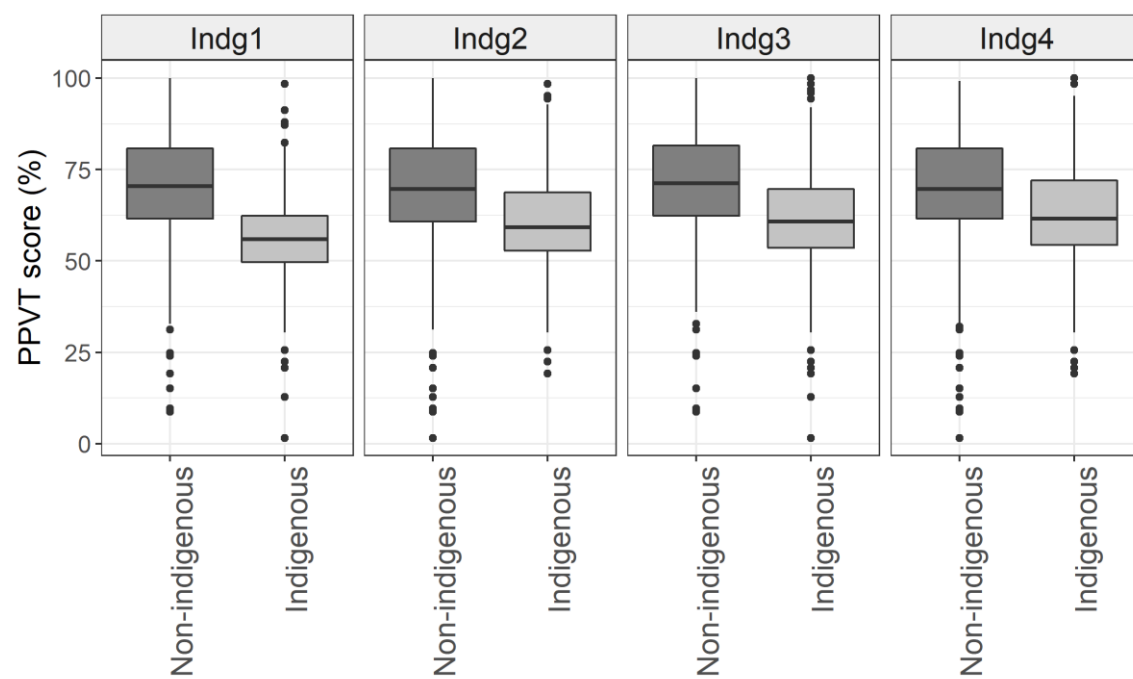
Figure 1 shows boxplots for the four binary indigeneity variables (*Indg1*, *Indg2*, *Indg3* and *Indg4*). It shows that the non-Indigenous children have higher PPVT scores than the Indigenous children regardless of indigeneity variable used. For example, the median PPVT score for non-Indigenous children by *Indg1*, *Indg2*, *Indg3* and *Indg4* definition is roughly 71, 70, 71 and 70 percent respectively, while the median score for Indigenous children equates to 56, 61, 62, 63 percent per definition. This shows that there is some variation in the difference between the Indigenous and non-Indigenous groups depending on the indigeneity variable used. While the score distribution of the non-Indigenous cases remains stable across indigeneity definitions (all have a median of around 70 percentage points), the position of the median scores for the Indigenous group, as well as of their upper and lower bounds (first and third quartiles), vary more by definition.

***The greatest difference between Indigenous and non-Indigenous children is observed for Indg1, the variable that uses the child's language. As shown in Source: Authors' creation***

Table 3, Indigenous children perform on average 15 PPVT percentage points lower than non-Indigenous children when child language is used to define their indigeneity status. The gap between Indigenous and non-Indigenous children average PPVT scores is smallest (6.8 percentage points) for *Indg4*, the variable that uses mothers' ethnicity to define indigeneity status of the cases. The difference between Indigenous and non-Indigenous groups is statistically significantly different in all instances (Source: Authors' creation

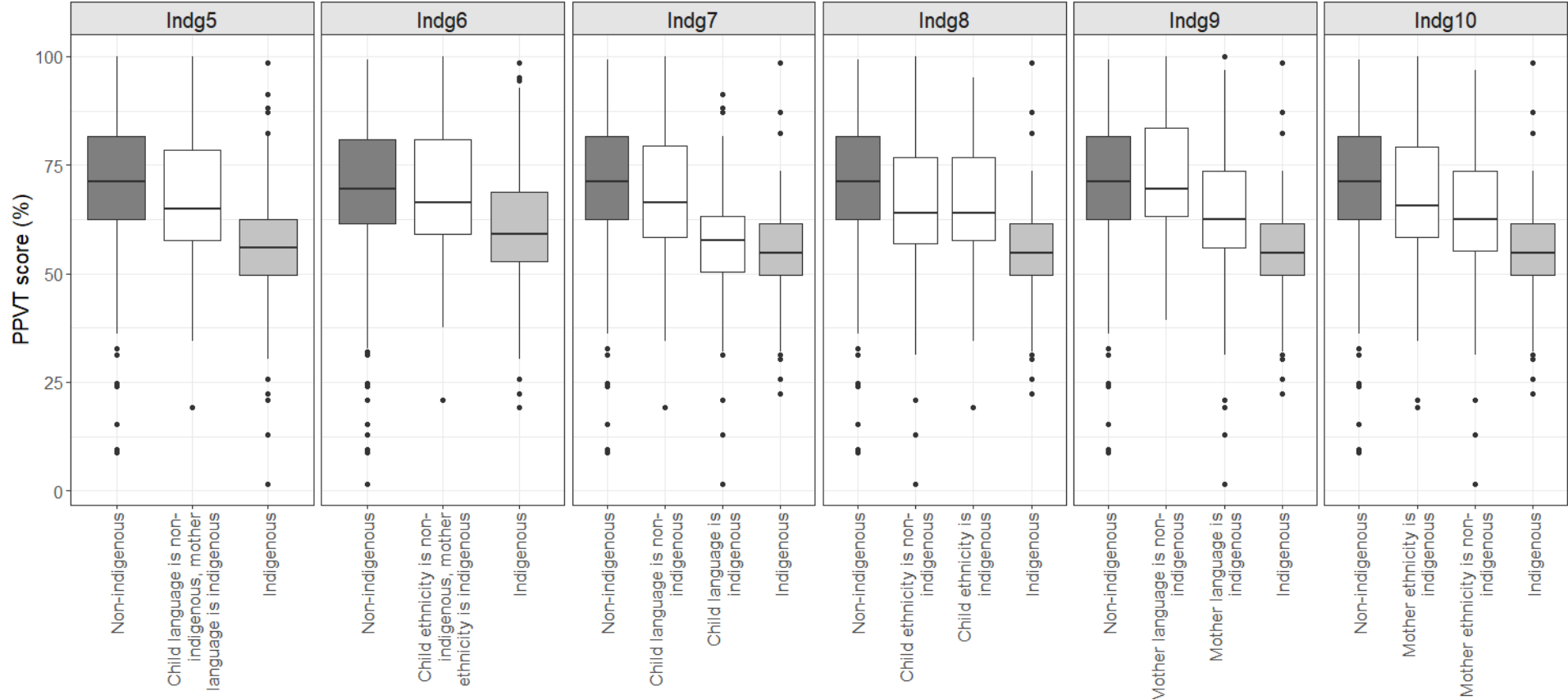
Table 3). These findings are in line with previous research.

***Figure 1. PPVT scores' distributions for binary variables***



Source: Authors' creation

Figure 2. PPVT scores' distributions for multi-category variables



Source: Authors' creation

### *Examining multi-category variables against PPVT scores*

Figure 2 presents boxplots for the multi-category indigeneity definitions. Variables *Indg5* uses language information to define indigeneity while *Indg6* uses ethnicity information (Table 2). Both variables have three categories. Looking at the position of the three boxes for these indigeneity variables we see that for *Indg5* (median percentage score: non-Indigenous 72, middle category 67, Indigenous 56) there is some more variation in the medians for each box/category, than for *Indg6* (median percentage score: non-Indigenous 70, middle category 68, Indigenous 61).

***This is further evidenced in the data presented in*** *Source: Authors' creation*

Table 3. When the 'Indigenous' group is compared to the 'non-Indigenous' group we see that for *Indg5* the difference is bigger (15.9 percentage points) than for *Indg6* (9.1 percentage points). Similarly, in the case of the middle categories, when these are compared to the non-Indigenous group, for *Indg5* children with a mother that speaks an Indigenous language (even if the child speaks Spanish) perform 4.2 PPVT percentage points lower than children categorised as non-Indigenous. The difference between these two groups is statistically significant. In contrast, for *Indg6*, children with a mother who is ethnically Indigenous (but the child is non-Indigenous) perform on average only 1.8 PPVT points lower than children in the non-Indigenous category and this difference is not statistically significant. Therefore, we see that similar to the binary variables, when language information is used to define Indigenous status greater differences in PPVT scores are seen between groups than when ethnicity information is used.

Figure 2 also presents the distributions for the four category variables (*Indg7*, *Indg8*, *Indg9* and *Indg10*). The Indigenous and non-Indigenous categories are the same for these four variables and are defined by combining language and ethnicity information from the child and the mother. For these variables, children classified as Indigenous have a median PPVT score of 55 percent, while non-Indigenous children a median score of 72 percent. In terms of average scores, Indigenous children perform on average 16.7 PPVT percentage points lower than the non-Indigenous children. This is the largest difference between Indigenous and non-Indigenous groups compared to the binary and three-category variables. Hence, we see that the more granular or exclusive the definition of indigeneity the greater the difference in the PPVT scores between the defined Indigenous and non-Indigenous groups.

Examining the middle categories once again shows that giving precedence to language information of the child results in larger differences between non-Indigenous and middle category pupils. In the case of *Indg7* the difference between the non-Indigenous category and the category 'child language is Indigenous' is 14.8 percentage points. When information on the language of the mother is used, we also see that children with a mother that speaks an Indigenous language on average perform lower than the non-Indigenous group, but the gap is smaller (7.3 percentage points, *Indg9*).

In contrast, when ethnicity information of the child or the mother are given precedence to establish the middle categories, results are less clear cut. In the case of variable *Indg8*, there is little difference in the scores of children in the middle categories. Average PPVT scores are similar for children categorised as ethnically Indigenous (on average they have 4.9 percentage points lower compared to the fully non-Indigenous group) and those who are ethnically non-Indigenous (5.9 percentage points lower than the fully non-Indigenous children). Interestingly, we see that the difference for ethnically Indigenous children is slightly smaller than that for ethnically non-Indigenous children when compared to the non-Indigenous group. Similarly, in the case of the middle categories for *Indg10* children in the category whose mother's ethnicity is Indigenous perform slightly better on average (4 percentage points lower compared to the non-Indigenous group) than those whose mother is ethnically non-Indigenous (7.6 percentage points lower to the non-Indigenous group).

The results for the differences in categories for *Indg8* and *Indg10* are worth highlighting as it shows that being ethnically Indigenous in some markers does not result in lower scores compared to being ethnically non-Indigenous. The dominant language in Peru is Spanish. Therefore, these results raise the question whether using language information to operationalise indigeneity provides a marker of experiencing exclusion rather than deficit of Indigenous children. These results may be highlighting the lack of processes to ensure that linguistic minorities can flourish and have equal opportunities in Peru.

## **Recommendations and research implications**

Following Indigenous Statistics and QuantCrit scholars, this paper established different operationalisations of indigeneity informed by a targeted review of empirical survey-based studies. This guided our analyses of different definitions of indigeneity using Young Lives Peru data. The review showed that there is limited thinking about the operationalisation of indigeneity in the examined quantitative research. Few studies discussed their rationale behind their chosen indigeneity operationalisation, and an even smaller number ran analyses for different operationalisations. When researchers do run analyses for different definitions of indigeneity, then most studies demonstrate that there are differences in outcomes depending on how indigeneity was operationalised.

To build on this finding, we provided an example using Young Lives Peru data to show that through few indicators, namely information about language and ethnicity, it is possible to operationalise indigeneity in ten different ways. Furthermore, by examining PPVT test scores we show that the chosen operationalisation matters: it matters as definitions using language show larger gaps between Indigenous and non-Indigenous children than the ones based on ethnicity, and it matters as some definitions show no gap between some groups. The use of language information, especially of the child, results in greater differences between the Indigenous and non-Indigenous groups, while being

ethnically Indigenous or having an Indigenous mother does not necessarily translate into having on average lower PPVT scores than ethnically non-Indigenous children or children with an ethnically non-Indigenous mother. Through our analyses we sought to highlight the myriad of ways indigeneity can be operationalised and that engaging in such an exercise can provide new insights. In this case, it can help better understand exclusion related issues.

Our results point to specific recommendations to consider the ways in which researchers engage with existing social surveys. First, given the multi-faceted nature of indigeneity, we recommend developing multi-dimensional indicators to provide more insightful markers of indigeneity. Second, when using existing datasets, the rationale behind chosen indicators should be clearly outlined. Third, when more than one indigeneity-relevant indicator is available, researchers should conduct sensitivity analyses using other indicators and provide comparisons. Fourth, through our review we identified survey-based studies in Peru such as CRISE and PERLA which contain a wide range of information on indigeneity – such as language, ethnicity, cultural self-identification and geography (Paredes, 2007; Telles, 2014). These are important quantitative surveys which should serve as useful references for how questions on indigeneity were developed.

Interestingly, most of the screened studies use or give precedence to language information to establish Indigenous and non-Indigenous groups. It is important to reflect on these results as researchers are using them to inform policy and to demonstrate that Indigenous groups are falling behind. While our findings using indigeneity variables based on language show a similar pattern, we are cautious about how to interpret them as we have not engaged with the measure of academic attainment in a way in which we can be certain that it is not influenced by biases from the dominant culture. Language information may be masking exclusion issues experienced by minority language groups (Walter & Andersen, 2013). Therefore, aspects around the cross-cultural validity of the tests used in these studies should be considered (Gillborn et al., 2018; Walter, 2010; Walter & Suina, 2019).

Basterra and colleagues (2011) point to the limits of using educational tests with language minority groups. Often these are ‘merely translations’ of their original versions and tend to be based on Euro-American realities. Similarly, for Treviño Villareal (2006) standardised educational tests not only measure the constructs intended (e.g. vocabulary, reading or math competences), but for Indigenous children entail measuring both educational competences and knowledge of the dominant culture. Therefore, these tests may provide information on aspects of the dominant culture that Indigenous children have learnt through formal schooling but are limited in providing information about the diverse ways of learning and knowledge of Indigenous cultures. Likewise, test results may in fact provide an indication of how well Indigenous children have mastered the dominant language, even if this is inadvertently done.

Young Lives documentation outlines the efforts made by the team to ensure the validity of its surveys and educational attainment tests across different contexts. Care and thought were put into developing their instruments and this article does not seek to criticise their work. Following Gillborn et al. (2018), Walters (2010) and Tuhiwai Smith (2013) we aim to increase awareness of how this data *has been* analysed, how it *can be* analysed and how best to *interpret* it. More broadly, we encourage those considering using results stemming from social survey research to critically reflect on reported findings. In the case of Young Lives, Cueto and León (2012), members of the Young Lives Peru team, warn against making comparative analyses within or between countries when using educational attainment data for groups that speak different languages. However, our review identified studies using Young Lives Peru data that ran analyses comparing test scores of groups speaking different languages. This raises questions about the validity of their drawn conclusions, as results could be masking biases within these tests rather than revealing accurate differences between groups. We do not contend that there are no differences between Indigenous and non-Indigenous children but highlight that it would be important to further understand the accuracy of our findings and of those reported in other studies. There is room for further research to understand the interpretative limits of the Young Lives data. For example, item discrimination analyses can be conducted to determine whether there are differences in how children in different groups perform for each PPVT item. To the best of our knowledge, this type of analyses have not been carried previously.

We are aware of the many factors that influence educational attainment. Relevant factors include socio-economic status, language of instruction, teacher quality and school infrastructure, among others (Ames, 2012; Arteaga & Glewwe, 2014; Hernandez-Zavala et al., 2006; Hynsjö & Damon, 2016; Skoufias et al., 2010; Van der Tuin & Verger, 2013). While exploring such confounders is beyond the scope of this study, these should also be considered when adopting QuantCrit and Indigenous Statistics lenses to research Indigenous educational issues. Overall, we argue that researchers should carefully consider how they operationalise and use indigeneity in quantitative analyses, as well as, consider the role of confounders. We recommend providing sensitivity analyses which can more deeply inform and explain the outcomes of their research.

## References

- Alcázar, L. (2019). *Brechas que perduran: Una radiografía de la exclusión social en Perú* / GRADE.  
BID. <http://www.grade.org.pe/publicaciones/brechas-que-perduran-una-radiografia-de-la-exclusion-social-en-peru/>
- Ames, P. (2012). Language, culture and identity in the transition to primary school: Challenges to indigenous children's rights to education in Peru. *International Journal of Educational Development*, 32(3), 454–462. <https://doi.org/10.1016/j.ijedudev.2011.11.006>
- Arteaga, I., & Glewwe, P. (2014). *Achievement gap between indigenous and non-indigenous children in Peru: An analysis of Young Lives survey data*. Young Lives.
- Basterra, M. del R., Solano Flores, G., & Trumbull, E. (2011). *Cultural validity in assessment: Addressing linguistic and cultural diversity* / edited by Maria del Rosario Basterra, Elise Trumbull, Guillermo Solano-Flores. (1st ed.). New York : Routledge, 2011.
- Cánepa, G. (2008). *The Fluidity of Ethnic Identities in Peru*. 46, 50.
- Castro, J. F., & Yamada, G. (2010). Las diferencias étnicas y de género en el acceso a la educación básica y superior en el Perú. *Centro de Investigación y Departamento de Economía de La Universidad Del Pacífico, Lima*.
- Cueto, S., Escobal, J., Penny, M., & Ames, P. (2011). Tracking disparities: Who gets left behind. *Initial Findings from PERU. Round 3 Survey Report*, 126.
- Cueto, S., Felipe, C., & Leon, J. (2019). *Venciendo la adversidad: Trayectorias educativas de estudiantes pobres en zonas rurales del Perú*. 80.
- Cueto, S., Guerrero, G., León, J., Zapata, M., & Freire, S. (2014). The relationship between socioeconomic status at age one, opportunities to learn and achievement in mathematics in fourth grade in Peru. *Oxford Review of Education*, 40(1), 50–72.
- Cueto, S., & León, J. (2012). *Psychometric Characteristics of Cognitive Development and Achievement Instruments in Round 3 of Young Lives*.  
<https://www.younglives.org.uk/content/psychometric-characteristics-cognitive-development-and-achievement-instruments-round-3-young>



- Cueto, S., Miranda, A., & Vásquez, M. C. (2016). Inequidades en la educación. *Grupo de Análisis para el Desarrollo*. <http://dspace.concytec.gob.pe:80/xmlui/handle/GRADE/140>
- Cueto, S., Singh, R., Woldehanna, T., Le Duc, T., Miranda, A., & Lives, Y. (2016). *Education trajectories in Ethiopia, India, Peru and Vietnam: From early childhood to early adulthood*.
- Dawes, A. (2020). *Measuring the Development of Cognitive Skills Across Time and Context: Reflections from Young Lives*. 13.
- Delprato, M. (2019). Parental education expectations and achievement for Indigenous students in Latin America: Evidence from TERCE learning survey. *International Journal of Educational Development*, 65, 10–25. <https://doi.org/10.1016/j.ijedudev.2018.12.004>
- Figueroa, A., & Barrón, M. (2005). *Inequality, Ethnicity and Social Disorder in Peru*. 77.
- Figueroa, R. (2003). Chikako Yamawaki. Estrategias de vida de los inmigrantes asiáticos en el Perú. *Hist&oacute;rica*, 27(2), 531–535.
- Flores-Bendezú, J., Calderón, J., Rojas, B., Alarcón-Matutti, E., & Gutiérrez, C. (2015). Desnutrición crónica y anemia en niños menores de 5 años de hogares indígenas del Perú: Análisis de la encuesta demográfica y de salud familiar 2013. *Anales de La Facultad de Medicina*, 76(2), 135–140. <https://doi.org/dx.doi.org/10.15381/anales.v76i2.11139>
- Gillborn, D., Warmington, P., & Demack, S. (2018). QuantCrit: Education, policy, ‘Big Data’ and principles for a critical race theory of statistics. *Race Ethnicity and Education*, 21(2), 158–179. <https://doi.org/10.1080/13613324.2017.1377417>
- Greene, S. (2016). Incas, Indios and Indigenism in Peru. *NACLA Report on the Americas*. <https://www-tandfonline-com.ezp.lib.cam.ac.uk/doi/abs/10.1080/10714839.2005.11724499>
- Guerrero, G., Leon, J., Rosales, E., Zapata, M., Freire, S., Saldarriaga, V., & Cueto, S. (2012). *Young Lives school survey in Peru: Design and initial findings*. Young Lives.
- Heckert, J. M. (2010). *Ethnic group disparities in academic achievement across four low-income countries*. <https://etda.libraries.psu.edu/catalog/10743>
- Hernandez-Zavala, M., Patrinos, H. A., & Sakellariou, C. (2006). *Quality of Schooling and Quality of Schools for Indigenous Students in Guatemala, Mexico and Peru*. World Bank Publications.

- Hospina, A. (2019, August 9). *Los Pueblos Indígenas en el Perú / El PNUD en Perú*. UNDP.  
<https://www.pe.undp.org/content/peru/es/home/presscenter/articles/2019/los-pueblos-indigenas-en-el-peru.html>
- Hutchinson, P., Dingwall, C., Kurtz, D., Evans, M., Jones, G., & Corbett, J. (2014). Maintaining the Integrity of Indigenous Knowledge; Sharing Metis Knowing Through Mixed Methods. *International Journal of Critical Indigenous Studies*, 7(1), 1–14.  
<https://doi.org/10.5204/ijcis.v7i1.118>
- Hynsjö, D., & Damon, A. (2016). Bilingual education in Peru: Evidence on how Quechua-medium education affects indigenous children's academic achievement. *Economics of Education Review*, 53, 116–132. <https://doi.org/10.1016/j.econedurev.2016.05.006>
- Ibarra, J. (2019). *Analysis of Peru's Education Policy for Internally Displaced Persons*. Instituto Nacional de Estadística e Informática. (2018). *La Autoidentificación Étnica: Población Indígena y Afroperuana*.
- International Work Group on Indigenous Affairs. (2021, March 18). *El Mundo Indígena 2021: Perú—IWGIA - International Work Group for Indigenous Affairs*.  
<https://www.iwgia.org/es/peru/4159-mi-2021-peru.html>
- Levi, J., & Maybury-Lewis, B. (2012). Becoming Indigenous Identity and Heterogeneity in a Global Movement. In *Indigenous Peoples, Poverty, and Development*. Cambridge University Press.
- Oré, B., Penny, M. E., & Madrid, S. (2012). *Selection and Induction of Supervisors for Fieldwork: Experiences from Young Lives in Peru* (YOUNG LIVES TECHNICAL NOTE NO. 24). Young Lives.
- Paredes, M. (2007). Fluid identities: Exploring ethnicity in Peru. *Centre for Research on Inequality, Human Security and Ethnicity*, 40, 1–35.
- Pasquier-Doumer, L., & Risso Brandon, F. (2015). Aspiration Failure: A Poverty Trap for Indigenous Children in Peru? *World Development*, 72, 208–223.  
<https://doi.org/10.1016/j.worlddev.2015.03.001>
- Sakellariou, C. (2008). Peer Effects and the Indigenous/Non-Indigenous Early Test-Score Gap in Peru. *Education Economics*, 16(4), 371–390.

- Skoufias, E., Lunde, T., & Patrinos, H. A. (2010). Social networks among indigenous peoples in Mexico. *Latin American Research Review*, 49–67.
- Snipp, M. (2016). What does data sovereignty imply: What does it look like? In *Indigenous data sovereignty Toward an agenda* (pp. 39–56). ANU Press.  
<https://library.oapen.org/bitstream/handle/20.500.12657/31875/624262.pdf?seq#page=63>
- Stein, M. T., & Lukasik, M. K. (2009). Chapter 79 - Developmental Screening and assessment: Infants, toddlers, and preschoolers. In W. B. Carey, A. C. Crocker, W. L. Coleman, E. R. Elias, & H. M. Feldman (Eds.), *Developmental-Behavioral Pediatrics (Fourth Edition)* (pp. 785–796). W.B. Saunders. <https://doi.org/10.1016/B978-1-4160-3370-7.00079-1>
- Telles, E. (2014). *Pigmentocracies: Ethnicity, Race, and Color in Latin America*. UNC Press Books.
- Treviño Villarreal, E. (2006). Evaluación del aprendizaje de los estudiantes indígenas en América Latina. Desafíos de medición e interpretación en contextos de diversidad cultural y desigualdad social. *Revista Mexicana de Investigación Educativa*, 11(28), 225–268.
- Trivelli, C. (2005). *Los hogares indígenas y la pobreza en el Perú: Una mirada a partir de la información cuantitativa*.
- Tuhiwai Smith, L. (2013). *Decolonizing methodologies: Research and indigenous peoples*. ZED BOOKS.
- UNICEF. (2010). *Estado de la Niñez Indígena en el Perú*. UNICEF, INEI.  
<https://repositorio.minedu.gob.pe/handle/20.500.12799/615>
- United Nations. (2004). *Martínez Cobo Study / United Nations For Indigenous Peoples*.  
<https://www.un.org/development/desa/indigenouspeoples/publications/2014/09/martinez-cobo-study/>
- Valdivia, M. (2007). *Etnicidad, antecedentes lingüísticos y la salud materno infantil en el Perú*. INEI. CIDE.
- Valdivia Vargas, N. (2011). *El uso de categorías étnico/raciales en censos y encuestas en el Perú: Balance y aportes para una discusión*. GRADE.
- Van der Tuin, M., & Verger, A. (2013). Evaluating teachers in Peru: Policy shortfalls and political implications. *GLOBAL MANAGERIAL EDUCATION REFORMS AND TEACHERS*, 127.

- Vásquez Huamán, E., Chumpitaz, A., & Jara, C. (2012). La inclusión de la niñez indígena: Repotenciando la educación intercultural bilingüe en el Perú. In *Perú sin Discriminación—Universidad del Pacífico*. <http://www.perusindiscriminacion.pe/publicaciones/la-inclusion-de-la-ninez-indigena-repotenciando-la-educacion-intercultural-bilingue-en-el-peru/>
- Walter, M. (2010). The Politics of the Data: How the Australian Statistical Indigene is Constructed. *International Journal of Critical Indigenous Studies*, 3(2), 45–56.  
<https://doi.org/10.5204/ijcis.v3i2.51>
- Walter, M., & Andersen, C. (2013). *Indigenous statistics: A quantitative research methodology / Maggie Walter and Chris Andersen*. Walnut Creek, CA : Left Coast Press, [2013].
- Walter, M., & Suina, M. (2019). Indigenous data, indigenous methodologies and indigenous data sovereignty. *International Journal of Social Research Methodology*, 22(3), 233–243.  
<https://doi.org/10.1080/13645579.2018.1531228>
- Williams, M., & Schertzer, R. (2019). Is Indigeneity like Ethnicity? Theorizing and Assessing Models of Indigenous Political Representation. *Canadian Journal of Political Science/Revue Canadienne de Science Politique*, 52(4), 677–696.  
<https://doi.org/10.1017/S0008423919000192>
- Yamada, G., Lizarzaburu, A., & Samanamud, K. (2011). *Diferencias étnicas en el mercado laboral peruano: Un estudio comparativo de brechas de ingreso*.  
<https://repositorio.up.edu.pe/handle/11354/366>
- Young Lives. (2017a). *A guide to Young Lives Research*.  
[https://www.younglives.org.uk/sites/www.younglives.org.uk/files/GuidetoYLResearch\\_0.pdf](https://www.younglives.org.uk/sites/www.younglives.org.uk/files/GuidetoYLResearch_0.pdf)
- Young Lives. (2017b). *A Guide to Young Lives Research—Section 5: Sampling*.  
<https://www.younglives.org.uk/sites/www.younglives.org.uk/files/GuidetoYLResearch-S10-Piloting.pdf>
- Young Lives. (2017c). *A Guide to Young Lives Research—Section 10: Piloting – Testing Instruments and Training Field Teams*.  
<https://www.younglives.org.uk/sites/www.younglives.org.uk/files/GuidetoYLResearch-S10-Piloting.pdf>

Zarate, A. (2011). Las representaciones sobre los indígenas en los libros de texto de Ciencias Sociales en el Perú. *Discurso & Sociedad*, 5(2), 333–375.

---

<sup>1</sup> The following combinations of search terms were used: “Young Lives study” AND “Peru” AND “Indigenous” OR “Native” OR “Ethnicity”; “Young Lives study” AND “Peru” AND “educational attainment” AND “Indigenous” OR “Ethnicity”; “Young Lives” AND “Indigenous pupils” “Peru” “educational attainment”; “Young Lives” AND “Indigenous pupils” AND “Peru” educational attainment; “Young Lives” AND Indigenous pupils in “Peru” educational attainment; Young Lives Indigenous pupils in Peru educational attainment. Searches were first conducted in English and later complemented by searches in Spanish using translations of these terms.

<sup>2</sup> We acknowledge the limitations around the data collection of PPVT in different languages. Any difference in PPVT may be the result of bias due to processes as opposed to due to children’s ability. Unfortunately, we are unable to examine further the limitations of the PPVT in our analyses in this article.

<sup>3</sup> Information on fieldwork operations in Peru available at <https://www.younglives.org.uk/content/selection-and-induction-supervisors-fieldwork-experiences-young-lives-peru>. Details on piloting operations are available in Section 10 at [https://www.younglives.org.uk/sites/www.younglives.org.uk/files/GuidetoYLResearch\\_0.pdf](https://www.younglives.org.uk/sites/www.younglives.org.uk/files/GuidetoYLResearch_0.pdf)

<sup>4</sup> It is worth noting that the Young Lives data set does not have a question on self-identification (e.g. worded around a sense of belonging), rather it asks the survey respondent (often the mother) what is their ethnicity and that of their child.

<sup>5</sup> The authors ran analyses with both definitions to check the robustness of their results based on their choice of indigeneity variable. However, they only reported results using the first definition and note that unless specified the results remain true for both definitions. Results using the restricted definition are available on request.

<sup>6</sup> As the goal of our analysis is to provide an example, information on the language and ethnicity of the mother and *not* of the father are used. This is informed by what was most common in the reviewed articles and for simplicity. Including information from the father as another marker would increase the complexity of our example.