Evaluating Measurement Invariance in the Measurement of Developmental Assets in Latino English Language Groups Across Developmental Stages

Okan Bulut¹, Jose Palma², Michael C. Rodriguez², and Luke Stanke³

Abstract
Noncognitive characteristics are gaining importance in addressing the persistent challenges facing youth in diverse settings. Measurement invariance of two youth developmental assets, Support and Positive Identity, is evaluated across grade levels and English language learner (ELL) subgroups of Latino students in 6th through 12th grade. Explanatory item response modeling is used to evaluate measurement invariance. The measurement of Latino students’ sense of support and positive identity varies depending on their developmental stage and language status. Students at later grade levels tend to require higher levels of Support to endorse items in the Support measure. There is a nonlinear relation between students’ grade level and item functioning for Positive Identity; students’ transition from middle to high school may influence the way they respond to Positive Identity items. This has implications for the measurement of assets with diverse Latino youth and for Latino youth development.

Keywords
measurement invariance, English language learners, validity, item response theory

Introduction
Social-emotional learning, developmental assets, and other so-called noncognitive characteristics are gaining presence in national dialogues addressing the persistent challenges facing youth. These are important components of 21st-century skills (Griffin, McGaw, & Care, 2012). The National Research Council (NRC; 2012) prominently identified noncognitive characteristics (interpersonal and intrapersonal skills) as core developmental characteristics to be successful in life and work. The NRC argued that cognitive, intrapersonal, and interpersonal skills can be taught and learned and contribute to successful adult outcomes. These skills also are found in the Common Core State Standards (Kyllonen, 2012) and include factors that shape school performance (Farrington et al., 2012), college and career readiness and success (Dymnicki, Sambolt, & Kidron, 2013), and more generally, positive youth development and resilience (Benson, Scales, Hamilton, & Sesma, 2006).

There exist many attempts to develop measures of such skills in diverse populations and at different ages or developmental stages, with significant measurement challenges (Griffin et al., 2012; Kyllonen, 2012). As a relatively new arena for measurement, rigorous measurement quality evaluation addressing the validity of score interpretations has been limited. “Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014, p. 11). In diverse settings, consistent interpretations must be permissible across groups to prevent misuse with marginalized populations and, more generally, with students facing persistent academic challenges. To establish a common interpretation framework in diverse settings, measurement invariance (MI; consistent score quality and meaning) should be confirmed across relevant subgroups. If invariance

¹University of Alberta, Edmonton, Canada
²University of Minnesota, Minneapolis, USA
³Minneapolis Public Schools, MN, USA

Corresponding Author:
Okan Bulut, Department of Educational Psychology, University of Alberta, 6-110A Education Centre North 11210 87 Ave. NW, Edmonton, Alberta, Canada T6G 2G5.
Email: bulut@ualberta.ca

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 3.0 License (http://creativecommons.org/licenses/by/3.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (http://www.uk.sagepub.com/aboutus/openaccess.htm).
holds, score differences can be assumed to reflect the differences in the construct being measured rather than irrelevant group differences due to measurement misspecifications (Millsap, 2010) or construct-irrelevant features (Haladyna & Downing, 2004). The investigation of MI is now a standard component of test development for most large-scale tests, particularly those with high stakes (AERA, APA, & NCME, 2014).

There is increasing interest in the assessment of noncognitive characteristics, 21st-century skills, and developmental assets among youth. These measures should be held to the same standards as educational and psychological tests if we are to defend their use to inform practice and policy. Developers of such measures must provide evidence of invariance across relevant subgroups, including race and ethnicity, language status, gender identity, and developmental stages—those groups for which the assessments are intended to be used (Millsap, 2010; Widaman, Ferrer, & Conger, 2010). The point is to identify subgroups that may have characteristics that potentially influence responses in construct-irrelevant ways. This is consistent with the standard approach to validation, based on a well-articulated interpretation and use argument (Kane, 2013).

In this article, MI is evaluated in a relatively new and flexible approach, through explanatory item response models (EIRMs). The study employs a measure of developmental assets administered to a large sample of middle and high schools’ Latino students with different English language proficiencies. This provides a rich context for the evaluation of MI and the exposition of these principles in the growing arena of social-emotional measurement.

The purpose of this study is to evaluate MI in measures of Support and Positive Identity. To present a coherent model of the process, the contexts of research in Latino youth development is presented in terms of the role of culture and language status; a review of positive youth development and developmental assets places the target measures in perspective; the evaluation of MI is reviewed; EIRMs to support this work are described; the methods and results of the study are summarized; and implications for measuring Support and Positive Identity are discussed. This serves two purposes, the study of MI of measures of Support and Positive Identity and the exposition of the evaluation of MI for measures of social-emotional characteristics for instructional purposes. In addition, as we have discovered, the evaluation of MI across developmental stages provides us with a deeper look into the developmental nature of developmental assets.

**Latinos and Youth Development**

U.S. Latinos are not only the largest ethnic minority group in the United States, 17% of the U.S. population in 2014, they are disproportionately young with 36% younger than the age of 20 (U.S. Census Bureau, 2014). Latino youth are more likely to face higher levels of developmental challenges including high poverty levels, residing in disadvantaged neighborhoods, attending low-quality schools, involvement in risky behaviors, and being raised by parents with low academic attainment and limited English skills (Eamon & Mulder, 2005; Kuperminc, Wilkins, Roche, & Alvarez-Jimenez, 2009). Latinos are diverse, based on generational status, country of origin, English language proficiency, age of immigration, residency status, and race (A. K. Fuligni & Perreira, 2009). Measures that assess development among Latino youth must consider the broad diversity of Latino youth populations to yield meaningful interpretations (Knight, Roosa, Calderón-Tena, & Gonzales, 2009). In addition, Latinos have been underrepresented and not equitably treated in developmental research, where studies that included Latino youth were deficit oriented and often unguided by a theoretical framework (Rodriguez & Morrobel, 2004). The need to address positive Latino youth development remains.

One of the core characteristics of Latino youth relevant to developmental considerations, particularly for school-age youth, is language and English language status (Rodriguez & Morrobel, 2004). In this study, the psychosocial role that English language acquisition plays among Latino youth is examined regarding its potential influence on the measurement of developmental assets. This recognizes an important characteristic contributing to the diversity of Latino youth, and one that may influence measurement in construct-irrelevant ways. English language fluency plays an integral part in the identity formation of Latino youth. Language skills determine how effectively they can interact within the spheres of English-speaking peers, teachers, and communities. Similarly, the amount of time in which Latino youth have contact with mainstream American culture affects how Latino youth develop identity and experience support in multiple contexts (Fuligni & Perreira, 2009). This multicontext influence on Latino youth is also recognized as a source of cultural resilience, as youth traverse and manage the multilingual and multicultural spheres of family, peers, teachers and school, and community (Taylor & Want, 2000).

**The Role of Language**

Second language acquisition has traditionally been studied in terms of cognitive development and educational contexts, but to a lesser extent regarding how it relates to social-emotional development. Language development plays an integral part in Latino youth development in the United States. In 2012, 74% of the Latino population reported to speak Spanish at home (U.S. Census Bureau, 2014). Maintaining a native language has been associated with retention of family values and cultural traditions as well as a protective factor against involvement in risky behaviors (i.e., an asset) among Latino youth (Kuperminc et al., 2009; Perez, Espinoza, Ramos, Coronado, & Cortes, 2009).
Limited English fluency has been associated with developmental risk factors including low socioeconomic status, residing in disadvantaged neighborhoods, low educational attainment by parents, and a perceived sense of stigmatization for not speaking English (Dawson & Williams, 2008; Garcia & Scribner, 2009). As Latino youth become more fluent in English, they also begin to interact more with the mainstream American culture and therefore be influenced by it (South, Crowder, & Chavez, 2005). Similarly, family responsibilities often increase for Latino youth as they begin to serve as mediators between their home culture and the surrounding communities when they act as family interpreters and translators (Morales, & Hanson, 2005; Weisskirch, 2005). The context facing recent immigrants is also challenging as it takes on average between 3 and 5 years to develop English-speaking fluency (Hakuta, Butler, & Witt, 2000). This makes English language status relevant in the study of Latino youth development and thus important in the evaluation of MI of developmental asset measures. However, language is just one component of culture and identity.

**Latino Cultural Context**

Not only are Latino youth underrepresented in developmental research, but also researchers who study Latino youth often treat Latinos as a homogeneous group, employ exploratory methods without theoretical grounding, and are largely deficit oriented (Rodriguez & Morrobel, 2004). A deficit orientation puts an emphasis on negative attributes and focuses on interventions addressing developmental risk factors. It is from the imbalance of assets and deficits that many of the problems faced by Latinos originate (Peck, Roeser, Zarrett, & Eccles, 2008; Rodriguez, & Morrobel, 2004). A call to refocus the orientation of Latino youth development research to a positive youth development approach has been raised by researchers (Kuperminc et al., 2009; Rodriguez, Morrobel, & Villarruel, 2003; Sesma, & Roehlkepartain, 2003). A positive youth development approach (Benson et al., 2006; Lerner, Almerigi, Theokas, & Lerner, 2005) aims to identify more appropriate environments where Latino youth can embrace their qualities and have a greater opportunity to enhance their abilities to thrive and succeed (Rodriguez et al., 2003).

Two developmental assets are of particular interest, Support and Positive Identity, because of the critical role they play in Latino youth development. Also, both constructs have strong developmental characteristics. Support is an external (interpersonal) asset as it characterizes engagement with others, whereas positive identity is an internal (intrapersonal) asset as it is manifested within the individual (Benson et al., 2006).

**Support**

The benefits of support received from family, community, and school among Latino youth have been well documented. Parental support has been associated with academic educational outcomes (Eamon & Mulder, 2005; Green, Rhodes, Hirsch, Suarez-Orozco, & Camic, 2008; Hill & Tyson, 2009; Ong, Phinney, & Dennis, 2006), as a protective factor against involvement in risky behaviors (Costello, Swendsen, Rose, & Dierker, 2008; Perez et al., 2009) and overall adjustment (Alvan, Belgrave, & Zea, 1996; Antrop-Gonzales, Velez, & Garrett, 2008). Similar benefits have been found when experiencing a supporting environment in the community (Faircloth, & Hamm, 2005; Urban, Lewin-Bizan, & Lerner, 2009), school (Alfaro, Umaña-Taylor, & Bamaca, 2006), and peers and other unrelated adults (Scales, Benson, & Mannes, 2006). However, research on the dynamics of Latino youth perceptions of support across developmental stages, particularly during adolescence, has been scarce (Grau, Azmitia, & Quattlebaum, 2009). There is some evidence that family support, community support, and school climate affect academic performance differently for 6th-, 9th-, and 12th-grade Mexican American students, indicating a potential interaction with developmental stage (Cabrera, Rodriguez, Palma, & Stanke, 2014). Support plays an important role in Latino youth development, but little is known about the role development plays in the perception of support.

**Positive Identity**

Identity development is a core milestone during adolescence (Erikson, 1968). For Latino youth, identity development must be considered with respect to ethnic identity development and consideration to the cultural and environmental contexts that influence identity formation. Ethnic identity has been considered part of the overall concept of personal identity (Phinney, 1996) and, for Latino youth, it is regarded with more centrality (Quintana & Scull, 2009). Identity development is complex. U.S. Latino youth must navigate between their culture of origin and the dominant or new culture where they reside as they work to form their identity (Quintana & Scull, 2009). At the same time, ethnic identity is highly influenced by language skills (often, unbalanced bilingual skills), age, interpersonal relationships with parents and peers (Phinney, Romero, Nava, & Huang, 2001; Quintana & Scull, 2009), and the broader society where youth develop their identity in the context of stigmatization and numerous adversities (Quintana & Scull, 2009), not to mention racism and anti-immigrant sentiments.

Positive ethnic identity has been associated with factors of well-being in Latino youth, including self-esteem (Umaña-Taylor, Gonzales-Backen, & Guimond, 2009), academic achievement (A. J. Fuligni, Witkow, & Garcia, 2005; Ong et al., 2006), and positive attitudes toward one’s social/cultural group and other groups (Whitehead, Ainsworth, Wittig,
& Gadino, 2009). However, Latino youth are more likely to live in disadvantaged neighborhoods with greater exposure to poverty, gangs, and crime, with limited support for developing a positive identity (Santrock, 2011). Like support, the role development plays in one’s sense of positive identity is less understood, compared with the role positive identity plays in youth development.

### Evaluation of MI

MI can be examined in many different ways. The challenge is to identify potential sources of measurement interference or differential functioning and hopefully address them. In this context, MI is based on evidence that a measure can be consistently interpreted across relevant subgroups, that there is no distortion in scores that is construct irrelevant and interferes with score interpretation. Score distortion can be caused by construct-irrelevant responses to items in a measure or test, including person characteristics that are associated with experiences or opportunities to learn (which may result in differential item functioning [DIF]; see, for example, Albano & Rodriguez, 2013), or shifts in item performance over time possibly due to changes in local culture or demographics, changes in curriculum or teaching methods, or major local or world tragedies or relevant events (which may result in item parameter drift [IPD]; Goldstein, 1983).

MI analysis is a class of methods appropriate for assessing invariance, which addresses the question of whether the instrument is measuring the same trait across all subgroups in a population or over measurement conditions. MI analysis is often implemented at the scale level via factor analytic methods (e.g., multigroup confirmatory factor analysis). Although a multigroup confirmatory factor analysis of subgroups may reveal equivalent factorial structures, item-level distortions can still be present (Zumbo, & Koh, 2005). Therefore, it is essential to conduct item-level analyses for evaluating item-level invariance among subgroups and to identify items that may affect score interpretation (Zumbo, 2007). In the context of Latino youth development with diverse Latino communities in terms of language status, we face multiple potential sources of measurement inconsistency, including differences in language status (DIF based on ELL status) and differences in developmental stages (IPD based on grade in school).

The IPD (Goldstein, 1983) is one approach to detecting measurement inconsistency at the item level. This study demonstrates IPD as an item-level method for investigating MI of youth development measures with a large Latino student sample, given English language learner (ELL) status, as currently receiving ELL services, received but exited ELL services, or never received ELL services, and grade levels as an indicator of developmental stages. EIRM (De Boeck & Wilson, 2004) was used to define various IPD models for detecting item-functioning variance as a result of ELL status and grade level. The findings of the study are discussed regarding invariance of youth development measures and possible implications for the substantive study of youth development itself.

Consistent score interpretation is a potential challenge given the nature of development and changes in response patterns to items in measures of youth characteristics. In standard measurement literature, violations of MI are thought of as construct irrelevant. As a concrete example, consider the measurement of school violence as perceived by students. We may measure perceptions of school violence through a number of items addressing various aspects of direct, indirect, physical, and relational violent behavior and incidents occurring in schools. Over time, because of the increasing level of attention to high-profile weapon use in a few cases across the United States and the dominance of zero-tolerance school policies, an item about the presence of weapons in school may function differently over time (i.e., IPD). As students are exposed to extreme examples of weapon use in schools, minor weapon infractions (e.g., carrying pocket knives or box cutters, etc.) appear to be less important or relevant or severe. If responses to the items in a measure of school violence should be a function of perceived school violence (the trait being measured), then we should be able to predict an individual’s responses to each item given his or her overall level of perceived school violence. Technically speaking, once we condition on overall level of perceived school violence, we might find that the item about the presence of weapons in schools functions differently over time. Even under the same level of perceived school violence, students would be less likely to indicate that the presence of weapons in schools is a problem. If the presence of weapons is an important indicator of school violence, and this aspect shifts in its functioning over time, the construct definition of school violence is changing, whether we change our interpretation of the scores or not—and unless we assess the presence of IPD, we have no reason to consider changes in score interpretation.

Goldstein (1983) introduced IPD as a way to measure changes over time in academic achievement exams. Item response theory (IRT) models are commonly used to investigate IPD under various contexts. An example includes a linear drift model of item location parameters (a measure of item difficulty) using a time-dependent item response model, where researchers argued that IPD can occur because of changes in curricular emphasis and be described as a function of time, that time-dependent item response models can describe data to maintain scales over long periods of assessment reporting (Bock, Muraki, & Pfeiffenberger, 1988).

There are many examples in the literature of IPD analyses. Researchers have investigated such tests as the Armed Services Vocational Aptitude Battery (ASVAB) and found little evidence of IPD, perhaps because aptitude is less variable than achievement over time (Chan, Drasgow, & Sawin, 1999). In an examination of measures of information literacy and global issues, items on the information literacy measure showed a greater magnitude of IPD, possibly resulting from rapid changes in the field (DeMars, 2004). In a recent examination...
of IPD in a measure of school climate, a few items became easier to endorse over time conditioned on overall school climate, suggesting that students may have become desensitized on some issues related to school climate as time passed, potentially changing the interpretation of school climate (Albano & Rodríguez, 2012).

From the body of research on IPD and exploratory studies investigating sources of IPD, we find potentially useful results. First, the construct being measured actually may change over time. Changes in the construct can indicate unintentional shifts in content over time (as items in a measure are replaced, reworded, or eliminated). Similarly, the construct as perceived by an individual may change due to real changes in cognitive abilities or social/emotional development. In addition, a construct as generally understood and defined by relevant populations may change over time because of the impact of individual-life or world events.

Assessing MI

In this study, MI is evaluated using EIRM. EIRM is a special case of IRT in the context of generalized linear mixed models (GLMMs) where it is possible to specify between-group differences in the latent constructs being measured (De Boeck & Wilson, 2004). EIRM extends the Rasch (1960) model, which describes the probability of responding in a particular way to an item as a function of the difference between the ability or trait level of person \( j \) and the location of item \( i \) on the trait continuum (commonly called item difficulty for achievement test items). This EIRM extension of the model can be written as a GLMM. The notation for this is a modified version of the hierarchical generalized linear modeling framework (Kamata, 2001). The Level-1 portion of the Rasch model, the item response level, can be written as a logistic regression model in terms of the log-odds of a correct response:

\[
\log \left( \frac{P_{ij}}{1-P_{ij}} \right) = \eta_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \cdots + \beta_{kj}X_{kij} = \beta_{0j} + \sum_{q=1}^{k} \beta_{qj}X_{qij},
\]

where \( \eta_{ij} \) is the log-odds (probability of correct response divided by probability of incorrect response) of person \( j \) endorsing the dichotomous item \( i \); \( \beta_{0j} \) is an intercept term and \( \beta_{ij} \) are coefficients associated with indicators \( X_{qj} \) through \( X_{kij} \) that indicate items. The indicators \( X_{qj} \) represent the \( q \)th dummy variable for person \( j \), and are coded as negative one when \( q = 1 \) and zero when \( q \neq i \). Because indicator coding uses negative one and zero, the item parameters can be interpreted as item location, the trait level associated with a 50% chance of endorsing the item. Level 2 of the model, the person-level, is

\[
\beta_{0j} = u_{0j} \quad \text{where} \quad u_{0j} \sim N(0, \tau_{j})
\]

\[
\beta_{1j} = \gamma_{10}
\]

\[
\vdots
\]

\[
\beta_{kj} = \gamma_{k0},
\]

where the intercept term, \( u_{0j} \), is a random effect for persons. Coefficients \( \beta_{1j} \) through \( \beta_{kj} \) are the fixed item locations (difficulties) \( \gamma_{10} \) through \( \gamma_{k0} \). Note that there is no fixed effect, \( \gamma_{00} \), included in the intercept term because the model does not consider a reference item when estimating the remaining items. When Level 1 and Level 2 are combined, the log-odds model is then

\[
\eta_{ij} = u_{0j} + \sum_{q=1}^{k} \gamma_{q0}X_{qij}.
\]

Because the indicator variables are coded as negative one when the response is associated with a particular item and zeros for all other items, the linear portion of the model can be simplified to \( \eta_{ij} = u_{0j} - \gamma_{q0} \), the difference between the person effect and the item location; another representation of the Rasch model. The Rasch model assumes that items possess the characteristic of local independence, meaning that after taking into account the parameters of the items and the persons, the responses to items are independent of one another. This implies that a single trait is responsible for each item response. In addition, MI is essential for ensuring generalizability across populations and test occasions (Rupp & Zumbo, 2006), supporting score interpretation and use; when MI does not hold, something other than the intended single trait is responsible for item responses.

EIRM and Study Design

As described earlier, we evaluated the role of two student characteristics, ELL status conceptualized as a potential source of DIF and grade level conceptualized as a potential source of parameter drift across developmental stages. This design employed student characteristics as features of the measurement model.

This study was conducted based on data gathered in an urban district in the state of Texas. Latino students who participated in the assessment are included in this secondary data analysis, including students in 6th through 12th grade. In addition, students were classified as (a) receiving ELL services, (b) received but exited ELL services, or (c) never received and/or never were eligible for ELL services. To receive ELL services, students report a home language other than English, are evaluated to determine their level of English proficiency, and receive a Broad English Ability score of 4 or lower on the Woodcock–Muñoz Language Survey–Revised. Latino students who have exited ELL services have satisfied
their ELL learning goals, performed at the advanced level (or lower with additional evidence of proficiency) on the Oregon English Language Proficiency Assessment, and have the ability to participate meaningfully in the general education program. Latino students who are not ELL students are those for whom English is their primary language.

The full explanatory model, including ELL status (two dummy variables to capture the three groups), grade level, and their interaction, is as follows:

\[
\begin{align*}
\beta_{0j} &= u_{0j}, \text{ where } u_{0j} &= N(0, \tau_j) \\
\beta_{ij} &= \gamma_{10} + \gamma_{11} (\text{ExitedELL})_j + \gamma_{12} (\text{ELL})_j + \gamma_{13} (\text{Grade})_j + \gamma_{14} (\text{ExitedELL} \times \text{Grade})_j + \gamma_{15} (\text{ELL} \times \text{Grade})_j + \gamma_{16} (\text{Grade} \times \text{Grade})_j + \cdots \\
\end{align*}
\]

In this model, \(u_{0j}\) is the trait level for person \(j\); \(\gamma_{q0}\) is the location for item \(q\) (1 to \(k\)) for a student belonging to the non-ELL group (the reference group) in sixth grade (Grade 6 is coded as zero, Grade 7 as 1, Grade 8 as 2, etc.; \(\gamma_{q1}\) is the item location difference for exited-ELL students; \(\gamma_{q2}\) is the item location difference for ELL students; \(\gamma_{q3}\) is the linear deviation in item location across grades for non-ELL students; \(\gamma_{q4}\) is the linear deviation difference in item location across grades for exited-ELL students; and \(\gamma_{q5}\) is the linear deviation difference in item location across grades for ELL students.

**Variations in EIRM Specification**

The EIRM, as a form of a GLMM, allows for a more flexible specification of the student characteristics in the model. As described above, ELL status is a factor (a categorical variable as in an ANOVA) and grade is a continuous linear predictor (a continuous variable as in regression). In the EIRM, grade is evaluated to the extent that its effect is linear across Grades 6 to 12, conditioned on other variables in the model. In other words, holding overall level of the trait constant (e.g., level of perceived Support), it evaluates the extent to which items shift in location across grades in a linear manner. This is referred to as the multigroup linear IPD model, where parameter drift may occur over grade or conceptually, developmental stages.

Alternatively, grade can be specified as a categorical variable, where the effect of each grade on item location is estimated independently. This is referred to here as the Grade \(\times\) Group factor model, where ELL status and grade are both considered factors (as in ANOVA) with interaction terms between each ELL status group and each grade level. The main advantage of the EIRM compared with other popular methods for testing MI (e.g., multigroup confirmatory factor analysis) is that the EIRM allows for the evaluation of IPD over grade levels not only linearly but also independently for each grade level (i.e., nonlinear drift).

**Method**

**Measures**

A questionnaire measuring developmental assets was administered to students in 6th through 12th grade in an English form and a bilingual form presenting English and Spanish versions on opposite pages. Students receiving ELL services were given the bilingual form and other students were given the option of which form to use. The Spanish version of the questions on the bilingual form were initially translated by one of the authors of this article and reviewed by a second author. The bilingual form was also reviewed by school personnel in the participating district with Spanish language backgrounds and a small number of issues were resolved through consensus. In part, the use of a bilingual form (presenting English and Spanish on opposite pages) was intended to avoid the influence of inappropriate translation. The translating authors and the school personnel reviewing these forms were of Mexican descent, similar to the vast majority of Latino students in the district.

The questionnaire included the Developmental Assets Profile” (DAP; Search Institute, 2005), which measures eight assets based on students’ experiences at the self, family, peer, school, and community levels. Data from two assets of interest, Positive Identity and Support, were used for the analysis. The Positive Identity measure consists of six rating-scale items, including issues related to feeling good about one’s self and future, dealing with disappointment, and having a sense of purpose. The Support measure consists of seven rating-scale items, including issues related to having parents who are encouraging and available to talk, having others in the community who are supportive, and being in a school that is supportive and encouraging.

Although the original items are on a 4-point scale, they were dichotomized to reduce the complexity of linear drift analysis in the EIRM, for illustration purposes only. Options “not at all or rarely” and “somewhat or sometimes” were coded as 0, and “very often” and “extremely or almost always” were coded as 1. In this way, item responses were then on a scale indicating more (1) or less (0) of the construct, as represented by each item. However, this is not recommended for general practice, but is indicated to simply facilitate ease of analysis and interpretation for our purposes here. Using the 4-point rating scale in the factor model evaluating nonlinear MI, for example, would produce more than 500 coefficients for the Support measure alone (three threshold coefficients are needed to describe the locations of the 4 rating-scale points, across 168 coefficients from the factor...
Table 1. Counts and Percentages of Latino Students by Grade Level and ELL Status.

| Grade | ELL n (%) | Exited-ELL n (%) | Non-ELL n (%) | Total |
|-------|-----------|------------------|---------------|-------|
| 6     | 1,621 (49.7) | 1,089 (33.3) | 554 (17.0) | 3,264 |
| 7     | 2,041 (44.9) | 1,775 (39.0) | 733 (16.1) | 4,549 |
| 8     | 1,595 (37.3) | 2,173 (50.8) | 508 (11.9) | 4,276 |
| 9     | 1,309 (33.4) | 2,279 (58.1) | 337 (8.5) | 3,925 |
| 10    | 961 (27.1) | 2,329 (65.7) | 253 (7.1) | 3,543 |
| 11    | 636 (24.2) | 1,809 (68.8) | 185 (7.0) | 2,630 |
| 12    | 562 (26.3) | 1,436 (67.3) | 137 (6.4) | 2,135 |
| All   | 8,725 (35.9) | 12,890 (53.0) | 2,707 (11.1) | 24,322 |

Note. ELL = English language learner.

model as described below). This would be cumbersome for the interpretation purposes here.

To use the EIRM framework in estimating person and item parameters (as in the Rasch model), the measures should meet the general assumptions of IRT, such as unidimensionality. Confirmatory factor analysis (CFA) was conducted in Mplus 6 (Muthén & Muthén, 2010) to confirm the unidimensional latent structure of the Positive Identity and Support measures. One challenge is that if MI does not hold, the CFA will indicate problems regarding the fit of a unidimensional model. Goodness-of-fit criteria, including root mean square error of approximation (RMSEA), Tucker–Lewis index (TLI), and comparative fit index (CFI), were used to evaluate model-data fit of the unidimensional CFA model for the two measures. CFI and TLI are incremental fit indices that range between 0.0 and 1.0 with values closer to 1.0 indicating good fit. RMSEA is an absolute fit index that is independent of sample size and thus performs well as an indicator of practical fit. For CFA models, Hu and Bentler (1999) suggested that for categorical data, RMSEA < .06, TLI > .90, and CFI > .90 indicate good fit. The CFA results indicated that Positive Identity (CFI = .94, TLI = .90, RMSEA = .095) and Support (CFI = .97, TLI = .95, RMSEA = .066) had acceptable levels of model-data fit on at least one criterion each; however, the RMSEA results were problematic. This is mostly an indicator that the simple unidimensional Rasch model may not be an appropriate approach to model item responses—which is consistent with the findings of measurement inconsistency reported below. The correlation between the two assets was .65, indicating a moderate relation between the two constructs (42% shared variance).

Sample Characteristics

As the focus of interest, the responses from 24,322 Latino students were analyzed. The counts of Latino students by grade and ELL status are reported in Table 1. The sample consisted of 36% ELL students, 53% exited-ELL students, and 11% non-ELL students. Approximately 50% were female, 5.9% were receiving special education services, 87.5% were identified by the district with a low socioeconomic status indicator, and 89% participated in free or reduced lunch.

Research Questions

We investigated one primary research question regarding MI: Does item functioning vary by grade, English language proficiency (language status), or their interaction? If drift does occur across grades as a function of ELL status, the interactions will be significant, requiring consideration of parameter variability across grade by ELL status. Other possibilities include the presence of drift that is constant across ELL status and the presence of DIF between ELL status groups. We simultaneously investigated MI as a function of both DIF by language status and IPD across grade level.

This led to substantive issues regarding the role of language status and developmental stage in the measurement of these two developmental assets. Important questions are addressed in the “Discussion” section regarding the impact of measurement variance on score interpretation for different subgroups of Latino students and regarding the developmental aspects of the measures themselves.

Data Analysis

In an exploratory study preceding this analysis (Stanke, Palma, Bulut, & Rodriguez, 2013), four models were fit to the data for Positive Identity and Support: (a) the Rasch model assuming MI, (b) a model to detect IPD treating ELL groups as invariant (no DIF), (c) a model to detect multiple-group IPD allowing for separate linear drift parameters for the three ELL groups (DIF × IPD), and (d) a model treating grade as a factor that interacts with ELL status that allows parameter drift across grades to be nonlinear and differ by subgroup (Grade × Group factor model). The models were fit to the Positive Identity and Support measures using the glmer function from the package lme4 (Bates, Maechler, & Bolker, 2012) in R (R Core Team, 2012). The final model, the Grade × Group factor model, appeared to fit best on a number of dimensions and resulted in item parameter estimates for each grade-by-group combination (Stanke et al., 2013). In any case, it was clear that MI was not a tenable assumption.

Based on these preliminary results, consistency of score interpretation of Positive Identity and Support is in part a function of Latino subgroup defined by language status, and development may not be represented on these two constructs in ways that are invariant; the measurement of Positive Identity and Support may vary in some items as a function of both grade level and language status. The results of the multiple-group linear IPD model are briefly presented and the linearity of the IPD is evaluated based on a graphical review of the Grade × Group factor model. The R code for estimating the Rasch and the multiple-group linear IPD models are provided in Appendix A.
### Table 2. Positive Identity Rasch Model Item Locations (Parameter Estimates), Standard Errors, and Item Fit Statistics.

| Item | Parameter Estimate | SE | Infit MSQ | Outfit MSQ |
|------|-------------------|----|-----------|------------|
| 1    | $\gamma_{10}$    | -2.06 | 0.02 | 0.99 | 1.02 |
| 2    | $\gamma_{20}$    | -2.36 | 0.02 | 0.96 | 0.97 |
| 3    | $\gamma_{30}$    | -2.45 | 0.02 | 0.83 | 0.67 |
| 4    | $\gamma_{40}$    | -0.36 | 0.02 | 1.02 | 1.08 |
| 5    | $\gamma_{50}$    | -1.41 | 0.02 | 0.90 | 0.88 |
| 6    | $\gamma_{60}$    | -2.54 | 0.02 | 0.90 | 0.87 |

Note. MSQ = mean square.

### Table 3. Support Rasch Model Item Locations (Parameter Estimates), Standard Errors, and Item Fit Statistics.

| Item | Parameter Estimate | SE | Infit MSQ | Outfit MSQ |
|------|-------------------|----|-----------|------------|
| 1    | $\gamma_{10}$    | -0.67 | 0.02 | 1.04 | 1.09 |
| 2    | $\gamma_{20}$    | -3.12 | 0.03 | 0.79 | 0.63 |
| 3    | $\gamma_{30}$    | -0.18 | 0.02 | 1.06 | 1.09 |
| 4    | $\gamma_{40}$    | -1.25 | 0.02 | 1.06 | 1.09 |
| 5    | $\gamma_{50}$    | -1.44 | 0.02 | 0.98 | 0.97 |
| 6    | $\gamma_{60}$    | -2.91 | 0.03 | 0.75 | 0.58 |
| 7    | $\gamma_{70}$    | -1.78 | 0.02 | 0.80 | 0.71 |

Note. MSQ = mean square.

### Results

To begin with, the Rasch model item locations and item fit statistics for Support and Positive Identity are reported in Tables 2 and 3. In the EIRM, the person parameter is normally distributed with a mean of zero; items are then located relative to the scale defined by person location. The infit and outfit fit indices indicate the extent to which the data matched the Rasch model (Bond & Fox, 2001), where values between 0.75 and 1.3 indicate adequate item fit (Smith, Schumacker, & Bush, 1998). Based on these suggested cutoff values, there is lack of adequate item fit for some items in both Support and Positive Identity, signaling that item difficulties may not be similar across ELL status and grade level. In the subsequent models, the items with adequate item fit were considered as anchor items in testing MI for Support and Positive Identity.

It is challenging to review a large number of coefficients; however, the results for the multiple-group linear IPD model are presented as the basis for evaluating MI. For Positive Identity, there are the person parameters (person trait level) and six items each with six coefficients; similarly for Support, there are seven items each with six coefficients. All coefficients are in the logit metric, due to the Rasch parameterization of the model.

Table 4 contains the 36-item coefficients for the multiple-group linear IPD model for the Positive Identity measure. Similarly, Table 5 contains the 42-item coefficients for the multiple-group linear IPD model for the Support measure. A Wald test assessed the statistical significance of each parameter being different from zero. The item locations, $\gamma_{q0}$ through $\gamma_{q6}$, were not tested relative to a null hypothesis; their locations are a function of the trait level required to be likely to endorse an item (whether that is different than zero is not of interest).

#### Positive Identity Multiple-Group Linear Drift

Regarding the Positive Identity measure, 5 of the 18 drift parameters from the multiple-groups IPD model, $\gamma_{q7}$ through $\gamma_{q12}$, had $p$ values less than .01; 3 of 6 drift parameters for exited-ELL students, 1 of 6 for ELL students, and 1 of 6 for existing-ELL students, and 1 of 6 for exiting-ELL students.
Table 5. Support Group and Linear Drift Model Item Locations (Parameter Estimates) and Standard Errors.

| Item | Parameter | Estimate | SE  | Description | Grade effect |
|------|-----------|----------|-----|-------------|--------------|
| 1    | $\gamma_{10}$ | -0.69    | 0.05| Exited-ELL  | Sixth-grade  |
|      | $\gamma_{11}$ | -0.72    | 0.08| Non-ELL     | baseline     |
|      | $\gamma_{12}$ | -1.00    | 0.05| ELL         |              |
|      | $\gamma_{13}$ | 0.06*    | 0.01| Exited-ELL  | Linear change|
|      | $\gamma_{14}$ | 0.16*    | 0.03| Non-ELL     |              |
|      | $\gamma_{15}$ | -0.00    | 0.02| ELL         |              |
| 2    | $\gamma_{20}$ | -3.81    | 0.09| Exited-ELL  | Sixth-grade  |
|      | $\gamma_{21}$ | -3.83    | 0.15| Non-ELL     | baseline     |
|      | $\gamma_{22}$ | -3.54    | 0.08| ELL         |              |
|      | $\gamma_{23}$ | 0.22*    | 0.02| Exited-ELL  | Linear change|
|      | $\gamma_{24}$ | 0.21*    | 0.05| Non-ELL     |              |
|      | $\gamma_{25}$ | 0.19*    | 0.03| ELL         |              |
| 3    | $\gamma_{30}$ | -0.77    | 0.05| Exited-ELL  | Sixth-grade  |
|      | $\gamma_{31}$ | -0.69    | 0.08| Non-ELL     | baseline     |
|      | $\gamma_{32}$ | -0.93    | 0.05| ELL         |              |
|      | $\gamma_{33}$ | 0.22*    | 0.01| Exited-ELL  | Linear change|
|      | $\gamma_{34}$ | 0.23*    | 0.03| Non-ELL     |              |
|      | $\gamma_{35}$ | 0.27*    | 0.02| ELL         |              |
| 4    | $\gamma_{40}$ | -1.69    | 0.05| Exited-ELL  | Sixth-grade  |
|      | $\gamma_{41}$ | -1.69    | 0.09| Non-ELL     | baseline     |
|      | $\gamma_{42}$ | -1.84    | 0.05| ELL         |              |
|      | $\gamma_{43}$ | 0.18*    | 0.02| Exited-ELL  | Linear change|
|      | $\gamma_{44}$ | 0.20*    | 0.03| Non-ELL     |              |
|      | $\gamma_{45}$ | 0.17*    | 0.02| ELL         |              |
| 5    | $\gamma_{50}$ | -1.60    | 0.06| Exited-ELL  | Sixth-grade  |
|      | $\gamma_{51}$ | -1.61    | 0.09| Non-ELL     | baseline     |
|      | $\gamma_{52}$ | -1.57    | 0.05| ELL         |              |
|      | $\gamma_{53}$ | 0.03    | 0.02| Exited-ELL  | Linear change|
|      | $\gamma_{54}$ | 0.02    | 0.03| Non-ELL     |              |
|      | $\gamma_{55}$ | 0.12*    | 0.02| ELL         |              |
| 6    | $\gamma_{60}$ | -3.52    | 0.08| Exited-ELL  | Sixth-grade  |
|      | $\gamma_{61}$ | -3.61    | 0.14| Non-ELL     | baseline     |
|      | $\gamma_{62}$ | -3.18    | 0.07| ELL         |              |
|      | $\gamma_{63}$ | 0.18*    | 0.02| Exited-ELL  | Linear change|
|      | $\gamma_{64}$ | 0.25*    | 0.05| Non-ELL     |              |
|      | $\gamma_{65}$ | 0.14*    | 0.02| ELL         |              |
| 7    | $\gamma_{70}$ | -2.14    | 0.06| Exited-ELL  | Sixth-grade  |
|      | $\gamma_{71}$ | -2.04    | 0.10| Non-ELL     | baseline     |
|      | $\gamma_{72}$ | -2.35    | 0.06| ELL         |              |
|      | $\gamma_{73}$ | 0.15*    | 0.02| Exited-ELL  | Linear change|
|      | $\gamma_{74}$ | 0.21*    | 0.03| Non-ELL     |              |
|      | $\gamma_{75}$ | 0.15*    | 0.02| ELL         |              |

Note. Item locations, $\gamma_q$ through $\gamma_q$, were not tested for significance. ELL = English language learner. *p < .01.

Regarding the Support measure, 18 of the 21 drift parameters from the multiple-groups IPD model had $p$ values less than .01. The significant drift estimates on the Support measure showed an interesting phenomenon: as grade increased, the likelihood of endorsing an item became more difficult—Positive coefficients indicate that endorsement requires higher trait levels across grade level. For five of the seven items, this effect was significant for all three ELL groups. This phenomenon that occurs to some extent across all seven items on the Support measure suggests that when item location parameters are treated as measurement invariant, 12th-grade students’ perceived level of Support is being underestimated and 6th-grade students’ perceived level of Support is being overestimated. In addition, DIF was observed as a function of Latino ELL subgroup.

Support Multiple-Group Linear Drift

To investigate the extent to which IPD is linear, the parameters for each item were plotted across grade level and a visual analysis of the linear drift parameters was compared with the grade-by-group factor model. Recall that the factor model employs dummy variables for each grade, considering grade to be categorical, and thus involves a much larger number of parameters. Because there are eight main effects (intercept, two ELL status dummy variables for the three ELL groups, and five grade dummy variables for the six grades) and 15 interaction effects ($3 \times 5$) for each item, the number of coefficients is too large to report in tables (138 for Positive Identity and 161 for Support). To facilitate interpretation of the results, graphic displays are presented for each item, illustrating the linear and factor model results. Figures 1 to 6 display the item parameters for Positive Identity; Figures 7 to 13 display the item parameters for Support.
Positive Identity Item Functioning

Based on the review of the graphic displays, three of the six Positive Identity items (Figures 1 to 3) illustrated linear drift parameters. Visual misfit (nonlinearity) occurred on three of the six Positive Identity items (Figures 4 to 6). The items illustrating nonlinear drift were about dealing with disappointment (Figure 4), dealing with things that are hard in life (Figure 5), and thinking about one’s purpose in life (Figure 6). Each graphic display is briefly interpreted to provide more concrete meaning regarding the results.

The first Positive Identity item (Figure 1) concerns having a sense of control in one’s life. The drift parameters were nonsignificant for this item, as seen in Table 4 ($\gamma_{13}$ to $\gamma_{15}$). Notice in Figure 1, the lines are relatively flat. The top solid line is for students receiving ELL services, the middle dashed line is for non-ELL students, and the bottom dotted line is for exited-ELL students. Also notice that the lines are relatively aligned with the location of the markers, indicating the item location independently for each group across grades from the factor model. The triangles (ELL) closely align to the top solid line, the squares (non-ELL) closely align to the middle dashed line, and the diamonds (exited-ELL) closely align to

Figure 2. Display of the parameter estimates for the Positive Identity item regarding feeling good about one’s self, by grade level for the linear IPD and Grade $\times$ Group factor models.

Note. IPD = item parameter drift; ELL = English language learner.

Figure 3. Display of the parameter estimates for the Positive Identity item regarding feeling good about one’s future, by grade level for the linear IPD and Grade $\times$ Group factor models.

Note. IPD = item parameter drift; ELL = English language learner.

Figure 4. Display of the parameter estimates for the Positive Identity item regarding dealing appropriately with disappointment, by grade level for the linear IPD and Grade $\times$ Group factor models.

Note. IPD = item parameter drift; ELL = English language learner.

Figure 5. Display of the parameter estimates for the Positive Identity item regarding dealing with life’s challenges, by grade level for the linear IPD and Grade $\times$ Group factor models.

Note. IPD = item parameter drift; ELL = English language learner.
the bottom dotted line. There appears to be no systematic drift in item location (level of Positive Identity required to have a sense of control) across grades.

However, notice there are differences in the location of these lines on the trait scale. This indicates significant DIF given ELL status. From Grade 6 to 12, the ELL student line is above those of the other two groups although the lines converge closer to Grade 12. To endorse this item regarding a sense of control in one’s life, students receiving ELL services must have a higher level of Positive Identity, indicated by the higher location of the ELL line on the trait level (Figure 1) and the higher item location estimate ($\gamma_{12}$, Table 4). If a constant item location was used to estimate students’ level of Positive Identity on this item, $-2.06$ as estimated by the Rasch model (Table 2), ELL students’ level of Positive Identity would be underestimated and that of exited-ELL and non-ELL students would be overestimated. Because endorsing this item requires a higher trait level for ELL students, endorsing the item should indicate a higher level of the trait for those students. The person estimation (scoring) model estimates a trait level that is commensurate with the location of the item for those individuals endorsing the item and reporting to have a sense of control in their life (and similarly for those who do not endorse the item, reporting to not have a sense of control). If the model does not correctly locate the...
item or uses the wrong item location, the person trait level is incorrectly estimated. Moreover, developmental profiles of differences in students across grades will be distorted (more on the meaning of this in the “Discussion” section).

The second and third Positive Identity items (Figures 2 and 3) follow similar linear drift patterns. Item 2 (feeling good about one’s self) is characterized by significant linear parameter drift for ELL students (the solid line in Figure 2 and the significant coefficient $\gamma_{25}$ in Table 4), where the trait level required to be likely to endorse the item (item location) decreases across Grades 6 to 12. Item 3 (feeling good about one’s future) is characterized by positive drift for exited-ELL and non-ELL students (Figure 3), where the item location increases across grades.

The fourth, fifth, and sixth Positive Identity items (Figures 4 to 6) are characterized by a common characteristic. First, the items function similarly for the three ELL groups, but the IPD across grades does not appear to be linear. For all three items, the item location increases from Grade 6 to 8, then decreases to Grade 12. This suggests that for all three groups, to endorse an item requires an increasingly higher level of Positive Identity to Grade 8, then a decreasing level of Positive Identity to Grade 12. This means that if a constant item location was used to estimate student Positive Identity,

![Figure 10](image1)
**Figure 10.** Display of the parameter estimates for the Support item regarding caring school, by grade level for the linear IPD and Grade × Group factor models.
Note. IPD = item parameter drift; ELL = English language learner.

![Figure 11](image2)
**Figure 11.** Display of the parameter estimates for the Support item regarding other adult support, by grade level for the linear IPD and Grade × Group factor models.
Note. IPD = item parameter drift; ELL = English language learner.

![Figure 12](image3)
**Figure 12.** Display of the parameter estimates for the Support item regarding loving supporting family, by grade level for the linear IPD and Grade × Group factor models.
Note. IPD = item parameter drift; ELL = English language learner.

![Figure 13](image4)
**Figure 13.** Display of the parameter estimates for the Support item regarding talking with parents, by grade level for the linear IPD and Grade × Group factor models.
Note. IPD = item parameter drift; ELL = English language learner.
students in Grade 8 would be underestimated and students in Grades 6 and 12 would be overestimated.

Support Item Functioning

The visual fit for the multiple-groups linear IPD model is better for the Support measure, as nearly all drift effects appear to be relatively linear. The visual analysis showed that the drift parameters for all three groups either followed linear drift for an item or the absence of drift (only 3 of the 21 parameters). Overall, linear parameter drift occurs in at least one group for every item, and all three groups for five of the seven items.

The first Support item (Figure 7) concerns asking parents for advice. The drift parameters were significant for exited-ELL and non-ELL students for this item, and nonsignificant for ELL students, as seen in Table 5 (γ₁₃ to γ₁₅). Notice in Figure 7 that there is some variation around the linear drift parameters (as can be seen by the markers), but the solid line (ELL) is flat, the dotted line (exited-ELL) is slightly increasing, and the dashed line (non-ELL) is more steeply increasing. There is another more subtle pattern that is similar to the pattern observed with the Positive Identity items: There is a slight increase for all groups from Grade 6 to 8 (or 9), then either a leveling off or slight decline to Grade 12. Again, this suggests that to be willing to endorse the item about asking parents for advice (that the student does ask parents for advice), it requires a higher level of Support overall for students in Grade 8 or 9 than other grades (more on the meaning of this in the “Discussion” section). However, the overall trend appears more linear than not, particularly for the remaining items.

Items 2, 3, and 4 (Figures 8 to 10) illustrate linear drift across grade, where all three ELL groups exhibit similar patterns, indicating no DIF across ELL subgroups. Item 5, regarding other adult support (Figure 11), illustrates linear drift for non-ELL students only. For Items 6 and 7 (Figures 12 and 13), all three groups are characterized by linear parameter drift, where the increase in item drift is more pronounced from Grades 6 to 8 and levels off to Grade 12.

Overall, on the Support measure, there is much less DIF across ELL subgroups, with the exception of Item 1 (Figure 7; asking parents for advice), where ELL students are more willing to endorse the item, holding overall Support constant. However, there is systematic, mostly linear, parameter drift across Grades 6 to 12. This occurs at some level with every item. The empirical item locations increase across grades, suggesting that the construct as a whole shifts across grades; students must have higher levels of Support overall to endorse each item (items are becoming more difficult to endorse across grades).

Discussion

Despite the importance of MI in terms of the validity of score interpretation and use, and the increasing attention to subgroup DIF (typically regarding gender or race), there is often a lack of concern for MI when an assessment is administered across different age groups or developmental stages, particularly with social/emotional domains. With increasing attention to meeting the developmental needs of diverse youth, closing achievement gaps, reducing school suspensions, and increasing graduation rates, assessments informing educators and policy makers on social/emotional characteristics of youth must be carefully examined to ensure appropriate, meaningful, and useful score interpretation across subgroups and developmental stages.

This study illustrates the use of EIRM to explore the evaluation of MI in the form of DIF (between ELL subgroups) and IPD (employing grade level as an indicator of developmental stage). DIF analysis has a long history and is a routine component in achievement test design and analysis (Livingston, 2006). IPD is typically used for evaluating MI across multiple administrations of a test over time. In this study, MI is evaluated by accounting for DIF and IPD. A questionnaire measuring developmental assets was administered to 6th- to 12th-grade students in an urban school district with a large Latino student population, including students receiving ELL services, students who had exited ELL services, and students who never received ELL services. Two assets, Positive Identity and Support, were evaluated for MI across ELL status and grade level.

The results of the MI evaluation indicated that some items function differently across ELL status and grade. For Positive Identity, the functioning of some items was influenced by grade level differentially based on language status. For Support, the functioning of nearly every item was influenced by grade level and, much less so, by language status. This has implications for interpretation consistency.

Implications for the Measurement of Positive Identity

The Positive Identity measure shows some DIF and IPD overall. However, this varied across items and language groups. Regarding DIF, for most items, ELL students needed a higher level of Positive Identity to have a likelihood of endorsing each item, and for some items, this was particularly true for 6th-grade ELL students, and less so for 12th-grade ELL students. For ELL students to be likely to agree with the components of Positive Identity (e.g., feeling in control of their life, feeling good about their future, dealing with life’s challenges and having purpose in life), they needed to have a higher level of Positive Identity overall. These items, based on a constant item parameter (assuming the same item parameter for everyone), are underestimating the Positive Identity of ELL students. This occurs when ELL students tend to endorse such items when they have higher levels of Positive Identity than non-ELL or exited-ELL students who endorse the items with lower levels of Positive Identity.

For three items, the parameter drift was nonlinear. For items about dealing with disappointments, dealing with hard
things in life, and thinking about one’s purpose in life, the drift appears to be systematic where the item location increases from 6th to 8th grade and decreases from 9th to 12th grades. This suggests that something triggers a shift in the increasing item location in 9th grade, causing the item location to then decrease. A potential reason could be that this is the time when students transition from middle to high school toward more independence. Students who endorse the items in later grades tend to have lower levels of Positive Identity overall; it becomes easier to agree with items in later grades and is most difficult to agree with items in 8th grade, controlling for overall Positive Identity.

Implications for the Measurement of Support

With the measure of Support, parameter drift was uniformly positive, where item locations increased from Grade 6 to 12 and mostly in a consistent way across ELL subgroups. All groups require lower levels of perceived Support to agree with items in the Support measure in earlier grades and higher levels of perceived Support at later grades. Because of parameter drift, estimates of perceived Support levels of students across grades are biased, where Support for students in younger grades is overestimated and Support in older grades is underestimated.

The questions remaining have to do with the impact of development on the measurement of developmental assets like Positive Identity and Support, and the variation that is found across Latino ELL subgroups. For Support (including family, school, and community support), it is more difficult to agree with statements about being supported for older students (requires an overall higher level of Support). So, interpretation of perceived Support across grade levels is distorted. When a 6th-grade student agrees with Support statements, his or her overall level of Support is estimated based on the location of the items in the measure. Because the item locations are lower on the trait scale in the linear drift model than in the constant-location (Rasch) model, their level of Support is overestimated. Similarly, because the item locations are higher on the trait scale for 12th-grade students, using a constant-location model underestimates Support for them. As students move from middle school grades to high school grades, Support becomes a slightly different construct (even with the same items), as the items become more difficult to endorse, indicating that if the same items are being endorsed in older grades, this should indicate a higher level of perceived Support than when endorsed in younger grades.

Implications for the Evaluation of MI

MI can be evaluated at the item level through DIF and IPD analyses. In this study, both DIF and a form of IPD (in cross-sectional data across grades) were applied in a single EIRM that allows the detection of both linear and nonlinear drift in item locations. Most items on the Positive Identity and Support measures examined here appeared to follow linear drift, but there were some items where drift was not linear and depended on language status (DIF). Methodologically, the use of IPD to measure drift across grade levels is not a common approach for examining parameter drift when data collection is cross-sectional, but this concept could be applied easily to a variety of cross-sectional settings. Through the use of the EIRM approach, both DIF and IPD and their interaction (IPD across grade levels depends on ELL status) can be evaluated succinctly. In addition, the model can be made flexible enough to assess the extent to which drift is linear or nonlinear.

Limitations and Future Research

Recommendations

The Latino student sample was obtained from several secondary schools in a single district. Although the proposed interpretations of IPD and DIF were not made for the purpose of generalizing across a broader Latino youth population, these specific findings may be a function of unique school district or school effects. However, a larger point of this work is to reinforce the practice of MI evaluation in noncognitive measures with diverse populations. The evaluation of MI should become common practice to defend interpretation and use of measures of social-emotional characteristics. Through wider examination of these issues, potential school effects or community-specific effects can be identified as well.

There are several areas recommended for future research. First, future studies could be completed to examine the patterns of drift across developmental stages for ELL, exited-ELL, and non-ELL students using different ethnic populations and different noncognitive measures. This also suggests that developmental stages be assessed more directly, rather than indirectly through grade level (as an indicator). This holds promise for contributing to developmental theory as well. Greater diversity in ethnic groups and measures will inform the future of this work and its implications. Second, the research could be extended to include predictors that explain the drift parameters. For instance, for items where drift varies as a function of ELL status, other information like immigrant status (especially regarding Positive Identity) or time spent interacting with adults (especially regarding Support) might serve explanatory functions. This type of study could be easily completed in the EIRM framework, as additional variables in the model. In this study, dichotomization of the items did not lead to a great deal of information loss at the scale level (see Appendix B). However, the models can be extended to the polytomous cases, taking advantage of the full information in more typical multipoint rating scales (although the researcher must be prepared to deal with hundreds of coefficients).

This study provides a unique example of examining MI of measures of developmental assets across developmental stages and language status. Although unconventional, given the typical examination of parameter drift in longitudinal contexts, the EIRM model provides a flexible and powerful approach to assessing IPD and DIF simultaneously. As the
effects of IPD and DIF and their interactions were estimated simultaneously, it also provided a more direct way to evaluate MI in the target measures. This also provided for a unique opportunity to interpret developmentally relevant youth characteristics (development as a major driver across the grade levels and ELL status) and their effects on the interpretation of scores. The purpose of this study was to investigate parameter invariance using a multiple-group linear drift model, where the findings exposed implications for interpreting youth development measures among ELL, non-ELL, and exited-ELL students across developmental stages.

### Appendix A

*Estimating the Rasch and IPD Models Using the lme4 Package in R*

#### Data Preparation

```r
### The data need to be reshaped from wide to long format
library(reshape2)
library(car)
support.long = melt(support, id=c("id", "ELL", "grade"), value.name = "response")

### Rename "variable", the default name given by the melt function, to "item"
support.long$item = support.long$variable

### Recode grade as a linear variable.
support.long$grade = recode(support.long$grade, "6=0; 7=1; 8=2; 9=3; 10=4; 11=5; 12=6")

### Create an interaction term for the factor IPD model
support.long$gr.ell.item = as.factor(paste(support.long$grade, support.long$item, support.long$ELL, sep = ""))
```

#### Model Estimation

```r
### Activate the lme4 package
library(lme4)

### Rasch model assuming measurement invariance across the groups
rasch = glmer(response ~ −1 + item + (1|id), family = binomial, data = support.long)

### Multiple-group linear IPD
linear.ipd = glmer(response ~ −1 + item:ELL + item:grade:ELL + (1|id), family = binomial, data = support.long)

### Group x Grade Factor IPD
factor.ipd = glmer(response ~ −1 + gr.ell.item + (1|id), family = binomial, data = support.long)
```

### Appendix B

*Figure B1.* Test information functions from the Support scale based on the polytomous items (left) and dichotomous items (right).
Acknowledgments

The authors thank Eugene Roehlkepartain, Peter Scales, and the data team of Search Institute for their support of this research and review of an earlier draft of the manuscript. We also greatly appreciate the feedback from Sage Open reviewers and action editor—Their comments were thoughtful and helpful.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research and/or authorship of this article.

References

Albano, A. D., & Rodriguez, M. C. (2012, April). Multilevel modeling of item parameter drift. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, British Columbia, Canada.

Albano, A. D., & Rodriguez, M. C. (2013). Examining differential math performance by gender and opportunity to learn. Educational and Psychological Measurement, 73, 836-856.

Alfaro, E. C., Umaña-Taylor, A. J., & Bamaca, M. Y. (2006). The influence of academic support on Latino adolescents’ academic motivation. National Council of Family Relations, 55, 276-291.

Alvan, S. L. J., Belgrave, F. Z., & Zea, M. C. (1996). Stress, social support, and college adjustment among Latino students. Cultural Diversity and Mental Health, 2, 193-203.

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

Antrop-Gonzales, R., Velez, W., & Garrett, T. (2008). Examining familial-based academic success factors in urban high school students: The case of Puerto Rican female high achievers. Marriage & Family Review, 43, 140-163.

Bates, D., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes (R package version 0.999999-0). Retrieved from http://CRAN.R-project.org/package=lme4

Benson, P. L., Scales, P. C., Hamilton, S. F., & Sesma, A. (2006). Positive youth development: Theory, research, and applications. In W. Damon & R. M. Lerner (Eds.), Handbook of child psychology: Vol. 1 (6th ed., pp. 894-941). New York, NY: John Wiley.

Bock, R. D., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. Journal of Educational Measurement, 25, 275-285.

Bond, T. G., & Fox, C. M. (2001). Applying the Rasch model (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Cabrera, J. C., Rodriguez, M. C., Palma, J. R., & Stanke, L. (2014, April). The influence of individual, family-related, and structural factors on Latino students’ academic performance. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Chan, K. Y., Drasgow, F., & Sawin, L. L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a cognitive ability test battery. Journal of Applied Psychology, 17, 610-619.

Costello, D. M., Swendsen, J., Rose, J. S., & Dierker, L. C. (2008). Risk and protective factors associated with trajectories of depressed mood from adolescence to early adulthood. Journal of Consulting and Clinical Psychology, 76, 173-183.

Dawson, B. A., & Williams, S. A. (2008). The impact of language status as an acculturative stressor on internalizing and externalizing behaviors among Latino/a children: A longitudinal analysis from school entry through third grade. Journal of Youth Adolescence, 37, 399-411.

De Boeck, P., & Wilson, M. (Eds.). (2004). Explanatory item response models: A generalized linear and nonlinear approach. New York, NY: Springer.

DeMars, C. E. (2004). Detection of item parameter drift over multiple test administrations. Applied Measurement in Education, 17, 265-300.

Figure B2. Test information functions from the Positive Identity scale based on the polytomous items (left) and dichotomous items (right).
Dymnicki, A., Sambolt, M., & Kidron, Y. (2013). Improving college and career readiness by incorporating social and emotional learning. Washington, DC: College & Career Readiness & Success Center, American Institutes for Research. Retrieved from http://www.ccrscenter.org/products-resources/improving-college-and-career-readiness-incorporating-social-and-emotional

Eamon, M. K., & Mulder, C. (2005). Predicting antisocial behavior among Latino young adolescents: An ecological systems analysis. American Journal of Orthopsychiatry, 75, 117-127.

Erikson, E. H. (1968). Identity, youth and crisis. New York, NY: Norton.

Faircloth, B. S., & Hamm, J. V. (2005). Sense of belonging among high school students representing 4 ethnic groups. Journal of Youth and Adolescence, 34, 293-309.

Farrington, C. A., Roderick, M., Allenworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., & Beechum, N. O. (2012). Teaching adolescents to become learners. The role of noncognitive factors in shaping school performance: A critical literature review. Chicago, IL: University of Chicago Consortium on Chicago School Research.

Fuligni, A. J., Witkow, M., & Garcia, C. (2005). Ethnic identity and academic adjustment of adolescents from Mexican, Chinese, and European backgrounds. Developmental Psychology, 41, 799-811.

Fuligni, A. K., & Perreira, K. M. (2009). Immigration and adaptation. In F. A. Villarruel, G. Carlo, J. M. Grau, M. Azmitia, N. J. Cabrera, & T. J. Chahin (Eds.), Handbook of U.S. Latino psychology: Developmental and community-based perspectives (pp. 99-113). Los Angeles, CA: SAGE.

García, E. E., & Scrineber, K. P. (2009). Latino pre-K education. In F. A. Villarruel, G. Carlo, J. M. Grau, M. Azmitia, N. J. Cabrera, & T. J. Chahin (Eds.), Handbook of U.S. Latino psychology: Developmental and community-based perspectives (pp. 99-113). Los Angeles, CA: SAGE.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. Journal of Educational Measurement, 20, 369-377.

Grau, J. F., Azmitia, M., & Quattlebaum, J. (2009). Latino families: Parenting, relational, and developmental processes. In F. A. Villarruel, G. Carlo, J. M. Grau, M. Azmitia, N. J. Cabrera, & T. J. Chahin (Eds.), Handbook of U.S. Latino psychology: Developmental and community-based perspectives (pp. 99-113). Los Angeles, CA: SAGE.

Green, G., Rhodes, J., Hirsch, A. H., Suarez-Orozco, C., & Camic, P. M. (2008). Supportive adult relationships and the academic engagement of Latin American immigrant youth. Journal of School Psychology, 46, 393-412.

Griffin, P., McGaw, B., & Care, E. (Eds.). (2012). Assessment and teaching of 21st century skills. Dordrecht, The Netherlands: Springer.

Hakuta, K., Butler, Y. G., & Witt, D. (2000). How long does it take English learners to attain proficiency? (Policy report). Santa Barbara: Linguistic Research Institute, University of California.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. Educational Measurement: Issues and Practice, 23, 17-27.

Hill, N. E., & Tyson, D. F. (2009). Parental involvement in middle school: A meta-analytic assessment of the strategies that promote achievement. Developmental Psychology, 45, 740-763.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6, 1-55.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. Journal of Educational Measurement, 38, 79-93.

Kane, M. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50, 1-73.

Knight, G. P., Roosa, M. W., Calderón-Tena, C. O., & Gonzales, N. A. (2009). Methodological issues in research on Latino populations. In F. A. Villarruel, G. Carlo, J. M. Grau, M. Azmitia, N. J. Cabrera, & T. J. Chahin (Eds.), Handbook of U.S. Latino psychology: Developmental and community-based perspectives (pp. 45-62). Los Angeles, CA: SAGE.

Kuperminc, G. P., Wilkins, N. J., Roche, C., & Alvarez-Jimenez, A. (2009). Risk, resilience, and positive development among Latino youth. In F. A. Villarruel, G. Carlo, J. M. Grau, M. Azmitia, N. J. Cabrera, & T. J. Chahin (Eds.), Handbook of U.S. Latino psychology: Developmental and community-based perspectives (pp. 213-233). Los Angeles, CA: SAGE.

Kyllonen, P. C. (2012). Measurement of 21st century skills in the Common Core State Standards. Princeton, NJ: K-12 Center, Educational Testing Service. Retrieved from http://www.k12center.org/rsc/pdf/session5-kyllonen-paper-tea2012.pdf

Lerner, R. M., Almerigi, J. B., Theokas, C., & Lerner, J. V. (2005). Positive youth development. Journal of Early Adolescence, 25, 10-16.

Livingston, S. A. (2006). Item analysis. In S. M. Downing & T. M. Haladyna (Eds.), Handbook of test development (pp. 421-441). Mahwah, NJ: Lawrence Erlbaum.

Millsap, R. E. (2010). Testing measurement invariance using item response theory in longitudinal data: An introduction. Child Development Perspectives, 4, 5-9.

Morales, A., & Hanson, W. E. (2005). Language brokering: An integrative review of the literature. Hispanic Journal of Behavioral Sciences, 27, 471-503.

Muthén, L. K., & Muthén, B. O. (2010). Mplus 6. Los Angeles, CA: Author.

National Research Council. (2012). Education for life and work: Developing transferable knowledge and skills in the 21st century. Washington, DC: The National Academies Press.

Ong, A. D., Phinney, J. S., & Dennis, J. (2006). Competence under conditions of language, parents, and peers in ethnic identity among adolescent Latino young adolescents: An ecological systems approach and ethnic identity in Latino college students. Journal of Adolescence, 29, 961-979.

Peck, S. C., Roeser, R. W., Zarrett, N., & Eccles, J. S. (2008). Exploring the roles of extracurricular activity quantity and quality in the educational resilience of vulnerable adolescents: Variable- and pattern-centered approaches. Journal of Social Issues, 64, 135-155.

Perez, W., Espinoza, R., Ramos, K., Coronado, H. M., & Cortes, R. (2009). Academic resilience among undocumented Latino students. Hispanic Journal of Behavioral Sciences, 31, 149-181.

Phinney, J. S. (1996). When we talk about American ethnic groups, what do we mean? American Psychologist, 51, 918-927.

Phinney, J. S., Romero, I., Nava, M., & Huang, D. (2001). The role of language, parents, and peers in ethnic identity among adolescents in immigrant families. Journal of Youth and Adolescence, 30, 135-153.

Quintana, S. M., & Scull, N. C. (2009). Latino ethnic identity. In F. A. Villarruel, G. Carlo, J. M. Grau, M. Azmitia, N. J. Cabrera, & T. J. Chahin (Eds.), Handbook of U.S. Latino psychology:
Developmental and community-based perspectives (pp. 81-98). Los Angeles, CA: SAGE.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danish Institute for Educational Research.

R Core Team. (2012). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rodriguez, M. C., & Morrobel, D. (2004). A review of Latino youth developmental research and a call for an asset orientation. Hispanic Journal of Behavioral Sciences, 26, 107-127.

Rodriguez, M. C., Morrobel, D., & Villarruel, F. A. (2003). Research realities and a vision of success for Latino youth development. In F. A. Villarruel, D. F. Perkins, L. M. Borden, & J. G. Keith (Eds.), Community youth development: Programs, policies, and practices (pp. 47-78). Thousand Oaks, CA: SAGE.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. Educational and Psychological Measurement, 66, 63-84.

Santrock, J. W. (2011). Educational psychology. New York, NY: McGraw-Hill.

Scales, P. C., Benson, P. L., & Mannes, M. (2006). The contribution to adolescent well-being made by nonfamily adults: An examination of developmental assets as contexts and processes. Journal of Community Psychology, 34, 401-413.

Search Institute. (2005). Developmental Assets Profile technical manual. Minneapolis, MN: Author.

Sesma, A., Jr., & Roehlkepartain, E. C. (2003). Unique strengths, shared strengths: Developmental assets among youth of color. Search Institute Insights & Evidence, 1(2). Retrieved from http://www.search-institute.org/research/insightsevidence/november-2003

Smith, R. M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. Journal of Outcome Measurement, 2, 66-78.

South, S. J., Crowder, K., & Chavez, E. (2005). Migration and spatial assimilation among U.S. Latinos: Classical versus segmented trajectories. Demography, 42, 497-521.

Stanke, L., Palma, J., Bulut, O., & Rodriguez, M. C. (2013, April). Investigating measurement invariance assumptions using item parameter drift across grade levels and ELL groups. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Taylor, R. D., & Want, M. C. (2000). Resilience across contexts: Family, work, culture, and community. Mahwah, NJ: Lawrence Erlbaum.

Umaha-Taylor, A. J., Gonzales-Backen, M. A., & Guimond, A. B. (2009). Latino adolescents' ethnic identity: Is there a developmental progression and does growth in ethnic identity predict growth in self-esteem? Child Development, 80, 391-405.

Urban, J. B., Lewin-Bizan, S., & Lerner, R. M. (2009). The role of neighborhood ecological assets and activity involvement in youth developmental outcomes: Differential impacts of asset poor and asset rich neighborhoods. Journal of Applied Developmental Psychology, 30, 601-614.

U.S. Census Bureau. (2014, September 8). Profile America: Facts for features. Hispanic Heritage Month 2014: Sept. 15-Oct. 15 (CB14-FF.22). Retrieved from http://www.census.gov/newsroom/facts-for-features/2014/cb14-ff22.html

Weisskirch, R. S. (2005). The relationship of language brokering to ethnic identity for Latino early adolescents. Hispanic Journal of Behavioral Sciences, 27, 286-299.

Whitehead, K. A., Ainsworth, A. T., Wittig, M. A., & Gadino, B. (2009). Implications of ethnic identity exploration and ethnic identity affirmation and belonging for intergroup attitudes among adolescents. Journal of Research on Adolescence, 19, 123-135.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structure equation models: Measuring the same construct across time. Child Development Perspectives, 4, 10-18.

Zumbo, B. D. (2007). Three generations of differential item functioning (DIF) analyses: Considering where it has been, where it is now, and where it is going. Language Assessment Quarterly, 4, 223-233.

Zumbo, B. D., & Koh, K. H. (2005). Manifestation of differences in item-level characteristics in scale-level measurement invariance tests of multi-group confirmatory factor analyses. Journal of Modern Applied Statistical Methods, 4, 275-282.

Author Biographies

Okan Bulut is an assistant professor of Measurement, Evaluation, and Cognition and a member of Centre for Research in Applied Measurement and Evaluation (CRAME) in the Faculty of Education at the University of Alberta, Canada. His current research interests include differential item functioning and measurement invariance in assessments, test reliability, Item Response Theory models, computerized adaptive testing, and technology-enhanced assessments.

Jose Palma is a doctoral student in Educational Psychology - Quantitative Methods in Education at the University of Minnesota. His research interests focus on addressing the methodological challenges in research and measurement of individuals from special and at-risk populations. Specifically, his work is centered on psychometric properties of assessments and socio-ecological contexts that influence human development.

Michael C. Rodriguez is the Campbell Leadership Chair in Education and Human Development and a professor of Quantitative Methods in Education at the University of Minnesota. His research interests include item writing and item response models, with substantive interests in early literacy and youth development. He consults internationally on item writing and test design.

Luke Stanke is a doctoral candidate in Educational Psychology - Quantitative Methods in Education at the University of Minnesota. He is also a member of an intra-university research group studying developmental outcomes of Minnesota's youth. His research is focused on analyzing item response data using parsimonious models.