Editorial

Assessing Children in Developmental Research

Challenges in Testing Through Infancy to Adolescence

Anke M. Weber[1], Samuel Greiff[1], and Dragos Iliescu[2,3]

[1] Cognition, Learning and Educational Assessment Research Group, Department of Behavioural and Cognitive Sciences, University of Luxembourg, Luxembourg
[2] Faculty of Psychology and Educational Sciences, University of Bucharest, Romania
[3] Department of Industrial Psychology, Stellenbosch University, South Africa

The need to make valid comparisons between groups is a challenge often faced in psychological research. In this, measurements across different groups are considered comparable (and, thus, group differences interpretable) if the administered test or questionnaire is shown to measure the same underlying construct with the same construct structure across the different groups (Greiff & Scherer, 2018; van de Schoot et al., 2012). Developmental psychology faces the additional challenge of measuring constructs across different age groups at various stages of development, which implies testing individuals at very different cognitive and non-cognitive stages. Developmental researchers attempt to come up with new and creative ways to assess developmental aspects (e.g., in very young children and infants); importantly for this editorial, for some of these developmental stages measures for which comparability between age groups has already been established sometimes cannot be used, or they simply do not exist.

To provide a specific example, one such construct is causal reasoning. Causal reasoning is assessed in many different ways across age groups. One procedure used for infants involves the "violation of expectation" paradigm: infants are introduced to some form of motion event, for example, an object rolling down a ramp. At its end, another object stands (Kotovsky & Baillargeon, 2000). Another procedure for the same construct involves infants reaching for toys after they observe causal relations (Sommerville et al., 2005). In children from 2 to 4 years, the blicket detector paradigm can be applied (Gopnik & Sobel, 2000; Gopnik et al., 2001; Griffiths et al., 2011): children are introduced to blocks that can be put on a machine, the "blicket detector". It works like this: (a) the experimenter proceeds to put one block at a time on the detector, which activates for one block but not the other, (b) the experimenter puts both blocks on the machine simultaneously and it activates, (c) the children are then asked to determine which of the blocks is a blicket and which is not. Again, for a different age group, in children from 4 to 6 years, gestures or children's explanations are frequently used to assess causal reasoning (Bonawitz et al., 2012; Pine et al., 2004).

The example of measuring causal reasoning shows that the same construct can be assessed in fundamentally different ways from gaze duration as an indicator for surprise in infants (violation of expectation paradigm, Baillargeon, 1994) to questionnaires and standardized tests in older children and adolescents (Gerstenberg & Tenenbaum, 2017). Given these special circumstances in developmental research, the question arises of how we can establish that the same construct with the same structure and content is measured across age groups and across measurement points (as individuals grow older). This question is especially important, since a lack of comparability – be it due to different methodological approaches or due to differences in underlying construct structure – across age groups can lead to inference problems because the results obtained in different groups are not directly comparable, and potential group differences are therefore not interpretable. Conclusions drawn from such results can be invalid and can, in the worst case, have negative consequences for children's development, for example, if an educational program that is beneficial for young children is rejected due to results in older age groups showing it to be non-effective (Chen, 2007).

This editorial addresses key challenges for measurement across age groups in developmental psychology, such as differences in assessment methods, narrow domains, small

sample sizes, lack of standardized assessment instruments for some constructs, and potential differences in construct structure across age groups.

Key Measurement Challenges Across Age Groups

Differences in Assessment Methods

The first challenge we identify is differences in assessment methods across age groups, since using the same assessment tool or even method is often impossible in developmental testing. This poses a problem especially in longitudinal studies that span multiple years – even though we note that those studies are the gold standard for understanding developmental trajectories. One example of how different assessment methods can lead to counter-intuitive results is studies on infants and young children suggesting that infants understand principles that preschool children apparently do not. This is the case, for instance, in the field of physical reasoning. Research on infants suggests that they reason intuitively about balance and support (Baillargeon, 1995), as measured by the violation of expectation paradigm. Since infants cannot speak about their reasoning on physical principles, the violation of expectation paradigm uses the method of gaze duration. The underlying assumption is that infants will gaze longer at a phenomenon that surprises them than at a phenomenon that they expect (Baillargeon, 1995). At the same time, research on somewhat older children, for instance, preschool children, usually applies a different set of methods, such as explicitly asking children to explain their reasoning (Bonawitz et al., 2012; Pine et al., 2004). Studies found that children struggle with physical reasoning during the preschool years, a task which infants seem to perform better at, which is a rather surprising finding. However, Krist and colleagues (2018) make the case that the methods applied in preschool are much more cognitively challenging than the violation of expectation paradigm and differences between infants and preschoolers might stem from this and not from actual differences in physical reasoning. In their study, Krist and his colleagues applied an eye-tracking paradigm with 2- to 6-year-old children and found that children became increasingly sensitive to physical principles such as the amount of support needed for objects to remain stable on a supporting surface. Their findings imply, contrary to the findings mentioned above that might have been confounded by different ways of measuring, development in physical reasoning.

This example shows that different ways of measuring the same construct can lead to different results. Therefore, the need to have comparable ways of measuring a construct, for example, with similar levels of cognitive demands, is one of the core challenges in establishing comparability through equivalence across age groups in developmental psychology (see Greiff & Iliescu, 2017).

Narrow Domains

The second challenge concerns very concrete and highly specific applications in real-world contexts and the measurement of constructs in a relatively narrow domain that assessments with young children often require. If broader domains are measured and the measure becomes too abstract, children might be unable to relate to the task, that is, to understand the questions or test stimuli (Arens et al., 2016), which leads to unreliable test results. For example, when measuring constructs related to the self, such as self-concept or self-efficacy, it is necessary to ask specific questions. In consequence, Arens and colleagues (2016) used highly specific questions to measure preschoolers' mathematics self-concept, for example, "Do you know lots of numbers?". Similarly, Oppermann and colleagues (2017) measured children's life science self-concept in a highly specific way as well, "Can you tell me how much you already know about plants?". The question arises whether results from these studies can be compared to broader and more abstract tests and questionnaires commonly used for adolescents or adults (Marsh et al., 2019), for example, "I have always believed that mathematics is one of my best subjects" for adolescents 15 years and older. It is likely that not only the structure but also the nature of the concepts changes as children and adolescents develop, which can also provide interesting insights into developmental trajectories. Therefore, in many cases, it remains unclear whether the reason for differences in mean, variance, or any other indicator of interest between age groups is an actual change in the underlying construct, or a change in the way the construct is measured, for example, from narrow and specific to broad and general. Differentiating between changes in the underlying construct and changes in measurement remains a challenge in developmental research.

Sample Sizes

The third challenge regards the relatively small sample sizes developmental researchers often work with, due to the nature of their experiments. Therefore, sample sizes larger than $N = 100$ children are scarce, limiting the possibilities to statistically check for equivalence across age groups, even if the assessment method is the same in these groups. For measures that can be implemented into larger surveys, for example, questionnaires or certain test instruments, (inter-)national surveys can enable researchers to investigate measurement invariance across age groups (Jones et al., 2016). Moreover, with samples of $N = 150$,

non-complex CFAs can be conducted to some extent and the CFI can be used as an index for goodness of fit (Chen, 2007).

In any case, it is clear that small sample sizes hamper statistical approaches to check for comparability between age groups and equivalence of measures in many studies; therefore, different approaches to establishing comparability need to be applied for developmental research.

Lack of Standardized Test Instruments

The fourth challenge we identify is a lack of standardized test instruments. The number of standardized test instruments for children of different ages is growing and many established batteries are now available to researchers (Dunn et al., 2015; Wechsler, 2012). However, often, and particularly for young children, standardized test instruments are still missing. Instead, researchers use measurement paradigms in which they develop their own assessment tasks, for example, violation of expectation, or the blicket detector (Baillargeon, 1995; Gopnik & Sobel, 2000; Sobel & Kirkham, 2006). Investigating these paradigms for comparability across age groups, for example with standardized measures that exist for older children and adults, would allow more detailed insights into construct structure and stability. For this, assessment methods for younger children could be adapted to older children, for example by using eye-tracking procedures (Krist et al., 2018), and could be compared to the results obtained in these older children with standardized tests, employing multi-modal testing. This approach could test whether the different procedures measure the same construct.

The challenge for developmental research coming along with this limitation is twofold: (1) standardized test instruments need to be developed with a focus on younger children, ensuring direct comparability with older age groups, and (2) research procedures for younger children and infants, such as the violation of expectation paradigm, need to be adapted for older children. This way, cognitive demands across age groups would be similar, and differences between age groups would be less likely to be caused by differences in cognitive demand.

Differences in Knowledge and Differences in Construct Structure

The fifth challenge concerns the need to distinguish between differences between age groups that stem from construct structure, prior knowledge, or the way the construct is measured in different age groups. Research suggests that the mechanisms underlying constructs, for example, reasoning, maybe the same for children and adolescents/adults and that the differences we find between age groups stem at least partly from higher knowledge that is present in adolescents and adults than in smaller children (Goswami, 2014). For example, a number of studies show that children tend to rely on perceptual similarities for reasoning, whereas adults rely on causalities. However, if children have knowledge about underlying causal relations or can infer them, they prefer causal reasoning over perceptual similarities as well, just as adolescents and adults do (Goddu et al., 2020; Griffiths et al., 2011). These differences in knowledge can potentially affect comparability between age groups and it is desirable to have similar prior knowledge conditions under which children and adolescents/adults are tested to ensure that differences in cognition are not merely differences in knowledge (that might then be mistaken for differences in the underlying processes).

Cognition and behavior can also serve different purposes and might have different meanings in different age groups (Putnick & Bornstein, 2016). For example, spatial skills are known to develop early in life and are even discussed to be innate by some researchers (Newcombe et al., 2013). However, certain subcomponents of spatial skills might present differently in different age groups and may undergo considerable development. Thus, the construct structure might differ and even change over relatively short time intervals. For example, navigation is one component of spatial skills that is not well developed in very young children (Immel et al., 2022; Newcombe et al., 2013). Therefore, if spatial skills are measured with a high reliance on navigation skills, younger children will inevitably do worse than older children, because navigation skills have not yet developed and therefore cannot represent spatial skills in young children. As a viable alternative, spatial skills could be measured with object representation, a core principle of spatial development (Baillargeon et al., 1985; Newcombe et al., 2013). But assuming the same underlying construct structure for spatial skills even while measuring them by relying on navigation will inevitably lead to mean differences because spatial skills present differently in younger children than in adolescents or adults. Consequently, this complicates the interpretation of differences between age groups. Mean differences could stem from either differences in construct structure from the way the construct is measured or from both. It is, with standard designs, almost impossible to disentangle the two. This shows that construct structure can play a role in the results obtained and has the potential to skew their interpretation. Therefore, the investigation of construct structure across age groups is of fundamental importance.

Finding (statistical) approaches to disentangle differences stemming from knowledge instead of change in the construct as well as measuring constructs in a way that the underlying structure is the same across age groups remains another challenge for developmental research.

Conclusions

We encourage developmental researchers to consider measurement aspects related to the challenges outlined in this editorial, for instance when comparing age groups in longitudinal as well as cross-sectional developmental studies. Missing the subtle nuances of comparability between measures can jeopardize the inferences drawn from studies and in the worst case have negative implications for children's lives. Challenges to comparability between age groups in developmental research can take the form of differences in assessment methods, specific issues in research with children, such as measuring constructs narrowly, or differences in knowledge between children and adults, as well as general threats to power, such as small sample sizes. We urge researchers to consider these challenges in their own work, particularly when submitting their work to *EJPA*.

In fact, we do notice that few if any of these challenges have been so far approached in papers published in *EJPA* and the assessment literature in general. Yet, they are fundamental assessment issues that provide researchers with important opportunities for creativity: not only these are problems that need to be solved, and any solution will have a rapid and noticeable impact, but solutions to these problems require creative approaches, and likely new methodological and statistical innovations. This is a seminal territory into which to delve to advance the field of psychological assessment. We encourage authors to submit papers dealing with comparability across age groups in general, and more specifically with the more unusual instances of (non-)comparability discussed in this Editorial.

References

Arens, A. K., Marsh, H. W., Craven, R. G., Yeung, A. S., Randhawa, E., & Hasselhorn, M. (2016). Math self-concept in preschool children: Structure, achievement relations, and generalizability across gender. *Early Childhood Research Quarterly, 36*, 391–403. https://doi.org/10.1016/j.ecresq.2015.12.024

Baillargeon, R. (1994). How do infants learn about the physical world? *Current Directions in Psychological Science, 3*(5), 133–140. https://doi.org/10.1111/1467-8721.ep10770614

Baillargeon, R. (1995). Physical reasoning in infancy. In M. S. Gazzaniga (Ed.), *A Bradford book. The cognitive neurosciences* (3rd ed., pp. 181–204). MIT Press.

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition, 20*(3), 191–208. https://doi.org/10.1016/0010-0277(85)90008-3

Bonawitz, E. B., van Schijndel, T. J. P., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology, 64*(4), 215–234. https://doi.org/10.1016/j.cogpsych.2011.12.002

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Dunn, L. M., Dunn, D. M., Lenhard, A., Lenhard, W., Segerer, R., & Suggate, S. (2015). *PPVT 4. Peabody Picture Vocabulary Test*. Pearson Assessment.

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. R. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 515–548). Oxford University Press.

Goddu, M. K., Lombrozo, T., & Gopnik, A. (2020). Transformations and transfer: Preschool children understand abstract relations and reason analogically in a causal task. *Child Development, 91*(6), 1898–1915. https://doi.org/10.1111/cdev.13412

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development, 71*(5), 1205–1222. https://doi.org/10.1111/1467-8624.00224

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37*(5), 620–629. https://doi.org/10.1037/0012-1649.37.5.620

Goswami, U. (2014). Inductive and deductive reasoning. In U. Goswami (Ed.), *Wiley-Blackwell handbooks of developmental psychology. The Wiley-Blackwell handbook of childhood cognitive development* (2nd ed., pp. 399–419). Wiley-Blackwell.

Greiff, S., & Iliescu, D. (2017). A test is much more than just the test itself. *European Journal of Psychological Assessment, 33*(3), 145–148. https://doi.org/10.1027/1015-5759/a000428

Greiff, S., & Scherer, R. (2018). Still comparing apples with oranges? Some thoughts on the principles and practices of measurement invariance testing. *European Journal of Psychological Assessment, 34*(3), 141–143. https://doi.org/10.1027/1015-5759/a000487

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2011). Bayes and blickets: Effects of knowledge on causal induction in children and adults. *Cognitive Science, 35*(8), 1407–1455. https://doi.org/10.1111/j.1551-6709.2011.01203.x

Immel, A.-S., Altgassen, M., Meyer, M., Endedijk, H. M., & Hunnius, S. (2022). Self-projection in early childhood: No evidence for a common underpinning of episodic memory, episodic future thinking, theory of mind, and spatial navigation. *Journal of Experimental Child Psychology, 223*, Article 105481. https://doi.org/10.1016/j.jecp.2022.105481

Jones, S. M., Zaslow, M., Darling-Churchill, K. E., & Halle, T. G. (2016). Assessing early childhood social and emotional development: Key conceptual and measurement issues. *Journal of Applied Developmental Psychology, 45*, 42–48. https://doi.org/10.1016/j.appdev.2016.02.008

Kotovsky, L., & Baillargeon, R. (2000). Reasoning about collisions involving inert objects in 7.5-month-old infants. *Developmental Science, 3*(3), 344–359. https://doi.org/10.1111/1467-7687.00129

Krist, H., Atlas, C., Fischer, H., & Wiese, C. (2018). Development of basic intuitions about physical support during early childhood: Evidence from a novel eye-tracking paradigm. *Quarterly Journal of Experimental Psychology, 71*(9), 1988–2004. https://doi.org/10.1177/1747021817737196

Marsh, H. W., van Zanden, B., Parker, P. D., Guo, J., Conigrave, J., & Seaton, M. (2019). Young women face disadvantage to enrollment in university STEM coursework regardless of prior achievement and attitudes. *American Educational Research Journal, 56*(5), 1629–1680. https://doi.org/10.3102/0002831218824111

Newcombe, N. S., Uttal, D. H., & Sauter, M. (2013). Spatial development. In P. D. Zelazo (Ed.), *The Oxford handbook of developmental psychology: Vol. 1: Body and mind* (pp. 564–590). Oxford University Press.

Oppermann, E., Brunner, M., Eccles, J. S., & Anders, Y. (2017). Uncovering young children's motivational beliefs about learning science. *Journal of Research in Science Teaching, 24*(2), 195–217. https://doi.org/10.1002/tea.21424

Pine, K. J., Lufkin, N., & Messer, D. (2004). More gestures than answers: Children learning about balance. *Developmental Psychology, 40*(6), 1059–1067. https://doi.org/10.1037/0012-1649.40.6.1059

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Sobel, D. M., & Kirkham, N. Z. (2006). Blickets and babies: The development of causal reasoning in toddlers and infants. *Developmental Psychology, 42*(6), 1103–1115. https://doi.org/10.1037/0012-1649.42.6.1103

Sommerville, J. A., Woodward, A. L., & Needham, A. (2005). Action experience alters 3-month-old infants' perception of others' actions. *Cognition, 96*(1), B1–11. https://doi.org/10.1016/j.cognition.2004.07.004

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Wechsler, D. (2012). *WPPSI-IV: Wechsler Preschool and Primary Scale of Intelligence – 4th edition*. Pearson Assessment.

Published online November 14, 2023

Anke M. Weber
Department of Behavioural and Cognitive Sciences
University of Luxembourg
2, avenue de l'Université
4365 Esch sur Alzette
Luxembourg
anke.weber@uni.lu

Samuel Greiff
Department of Behavioural and Cognitive Sciences
University of Luxembourg
2, avenue de l'Université
4365 Esch sur Alzette
Luxembourg
samuel.greiff@uni.lu

Dragos Iliescu
Faculty of Psychology and Educational Sciences
University of Bucharest
Sos. Panduri 90
050657 Bucharest
Romania
dragos.iliescu@fpse.unibuc.ro

