nature portfolio

Peer Review File

Contributions of Spontaneous Discrimination and Conceptual Knowledge to Children's Understanding of Emotion

Corresponding Author: Dr Wanze Xie

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This paper reports on relations between several facets of emotion perception in youth. The authors use EEG to test whether children discriminate facial expressions of emotion and other tasks to assess conceptual knowledge of emotion words and attributions of emotion expressions to scenes. The authors find that differences in positive vs negative faces are detected early, and more fine-grained differences emerge with age. They call this a measure of perceptual discrimination. They also collected measures of conceptual knowledge (semantic similarity ratings), sorting (connecting facial expressions to 'causes'), and matching (dragging mouse to connect two facial expressions of the same emotion). Using RSA, they show that sorting performance is related to perceptual discrimination at young but not older ages. Sorting performance at all ages is related to conceptual knowledge. Conversely, matching performance is predicted by conceptual knowledge only at older ages and perceptual discrimination at all ages.

This is an interesting paper that has several strengths. First, it uses a battery of tasks to probe several components of emotion development. Second, it tackles these questions in China, adding to the diversity of our typical participant pools. Third, it uses both brain and behavioral data. My questions and concerns with the paper are listed below:

Major concerns
- I had the most questions about the conceptual knowledge task. (i) In the main text, it makes it seem like the ratings were between two emotion words (e.g., anger and sadness), but in the methods the task is described as rating similarity between emotion words and other relevant terms (e.g., anger and crying). Which is it, as these seem like very important differences. (ii) The task seems to be analyzed using a series of correlation coefficients which are then averaged together. It seems like the more conventional way to analyze a task like this (Barrett, 2004, JPSP; Nook et al., 2014, NHB; Suvak et al., 2011, Abnormal). It's strange that negative emotions would be seen as more similar as age increases, so I'm curious if analyses using an MDS approach provides a different set of results or the same. (iii) This result of increased similarity across age for negative emotions also seems to emerge only in a correlation where age is rounded to the nearest integer (SI Fig 5). Can you use an unrounded age estimate for continuous age analyses? (e.g., if someone is 5 years, 6 months, and 15 days old, they're 5.55 years old). (iv) In several places, this task is described as assessing "norms" about emotions. Although I'm completely on board with the assumption that kids' conceptual knowledge of emotions is learned from the social milieux, I don't think a given individual's semantic understanding of emotions is a direct measure of a group's norms. There can be deviance between what one person thinks and what a group tends to think, no? I'd recommend clarifying or removing these points.
- I was also left unsure of whether young children did or did not seem to perceptually discriminate negative emotions according to EEG data. The analyses clearly show that pos/neg faces are discriminated and that neg discrimination increases with age. But are (for example) oscillating sad and angry faces producing an oddball effect in 5yos? I'm thinking that line 208 is saying this, but I'm not certain. This seems important theoretically, given the arguments that emotion perception may follow a "broad to narrow" developmental trajectory. It should also lead the authors to revise the comment that "discrimination was firmly in place by age 5" (which is made in the abstract and throughout the paper), as this is only true when faces differ by valence.
- How many times did the following occur: "The experimenter would skip a particular item if the child still found an item incomprehensible after the explanation" across each age? It's important to note exclusion rates.

Smaller concerns

- The abstract is quite jargony. It's not clear how discrimination and conceptual knowledge are operationalized, making it difficult for readers to get a sense of what was done and what was found.
- Terminology was difficult throughout, actually. MT, sorting, conceptual knowledge, etc. are fairly abstract and unintuitive terms, and I had to keep going back to the methods section to remind myself what each was representing. I'd recommend clearly laying out each construct you're interested in and naming how you measure each. Then it likely best to refer to both throughout. For example, "We measured conceptual knowledge about emotions using the semantic similarities task… Conceptual knowledge (i.e., arousal focus extracted from the semantic similarities task) was associated with xyz".
- If these "studies" are all on the same children, are they different studies? It seems to me that this is one study with several tasks.
- The intro follows a strange structure in that the study is introduced several times. Typically, the intro reviews each section of background literature, culminating in an open research question. This is then used to introduce the study just once at the end as a way to answer this question.
- In intro, sentence 2 is written in a way that makes it seem like it's in tension with sentences 3-4 but aren't they compatible?
- In this sentence, the terminology is not clear "due to the processing of structural changes between facial configurations rather than featural changes". What exactly is meant by facial configurations vs featural changes? The inverted oddball control is actually extremely important for the logic of this study.
- Beware causal language throughout, as the study is correlational (e.g., "conceptual knowledge significantly influenced the results" in SI)
- Line 77 – are they really "experiments"?
- Abstract N should be final usable sample, not the sample you collected (30 excluded)

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
In this paper, the authors report a series of 3 studies of emotion perception and emotion concept knowledge in 5- to 10-year-old children. Study 1 examined a neural correlate of (implicit/automatic) perceptual discrimination between stereotypical facial configurations of four emotion categories (via FPVS); Study 2 examined children's conceptual knowledge of these categories in terms of similarities in associations with common behavioral and contextual features (e.g., crying); Study 3 used both sorting and matching tasks to assess how children judge stereotypical facial configurations for emotion. The authors find distinct trajectories for perceptual discrimination and conceptual knowledge, and that their relationships to emotion judgment change over developmental time.

I found this an elegant set of studies that nicely builds on constructionist theories and recent empirical work (namely Brooks & Freeman, 2018) to illustrate how domain-general processes shape emotional development in different but intersecting ways, as children move through different ways of interfacing with the world. 'Emotional' input (that is, inputs that scaffold interactions with others) is at first largely perceptual (e.g., based on facial movements), and becomes more conceptual (e.g., anchored in language) over time. With this in mind, I think there are a few ways the authors could make the value of their contribution more evident and accessible to a broader set of readers.

1. More definition and explanation of the 'perceptual' vs 'conceptual' distinction would help clarify the nature of the processes at play.

a. Intuitively, it would seem that perceptual discrimination happens automatically, without conscious awareness – whereas conceptual discrimination is a deliberate classification. Is that a reasonable statement, or could the oddball responses also be based in part on conceptual knowledge? Such a possibility would be consistent with studies showing that the lexical categories modulate early neural responses in oddball paradigms (e.g., Boutonnet et al., 2013).

b. In Study 1, the authors confirm that faces were being processed configurally, as inverted faces elicited lower oddball effects. How, if at all, does the distinction between featural and configural processing map onto the distinction between perceptual and conceptual discrimination? Does configural processing imply the 'concept' of a face?

c. There are also a few points where it is implied that broad, valence-based distinctions between facial configurations are 'perceptual' distinctions – or, at least, that they are not yet sophisticated enough to count as 'conceptual'. This implication does not align with the thinking in many of the constructionist papers cited here. Instead, valence is conceived of a broad conceptual dimension that children may leverage to distinguish between faces, based presumably on their embodied experiences of encountering these faces in the interactions with caretakers and others around them.

2. Can the authors say more about why they selected the four emotion categories that they did (anger, fear, happiness, and sadness), what implications this has for the present work? I am curious, for example, why only one positive emotion category was included, and why categories that (at least in some cultural groups) are known to develop later on (e.g., disgust, shame) were not. Relatedly, do these four categories map onto this identified by Jack and colleagues (2016)?

3. I wasn't quite sure how to interpret the results of Study 1. It is stated both that "some emotion categories were better discriminated with increasing age" (p 7) and that "there were no significant differences among the three child groups" (p 8). There is also a reference to behavioral findings being reflected in EEG activity (p 7); what are the behavioral findings being

referenced – are they from this study?

4. The interpretation of Study 3 (and the paper overall) is that perceptual discrimination influences emotion judgment earlier on in childhood, but that conceptual knowledge comes online to dominate later in childhood. The actual results, however, are more complex than this. In the sorting task, there is no interaction between conceptual knowledge and age (only a positive main effect). In the matching task, there is no interaction between perceptual discrimination and age (only a positive main effect). So it seems an alternative explanation is that these tasks differentially tap into the proposed underlying mechanisms – that there is no cohesive intersection of the mechanisms that extends across all contexts for emotion judgment. I'd like to see the authors engage with these nuances in their discussion.

5. This work inspires a lot of connections with the existing literature, both developmental and otherwise. To the extent possible within this submission format, the authors could consider integrating the following references:
a. "cultural frameworks for displaying emotions" (Keller & Otto, 2009; Wang, 2001)
b. "linguistic input about emotions" (Borghi et al., 2017; Hoemann et al., 2019; Nook et al., 2017; Vigliocco et al., 2014)
c. "evidence that conceptual knowledge guides children's reasoning about emotion" (Ogren & Sandhofer, 2022; Price et al., 2022)
d. "older children will show more distinct neural responses between configurations" (Gao et al., 2024)
e. "[children] conceptualize various negative emotions as more similar to each other with age" (Nook et al., 2018)

Minor miscellaneous points:
6. Please briefly define "valence-based judgement" (p 3).

7. Given what we know about emotional development across the lifespan, I might say that children development progresses to *more* nuanced negative emotion differentiation (p 3) – in other words, there are still levels of nuance yet to come!

8. A typo in line 77 states that 4 experiments were conducted (instead of 3).

9. Can the authors say just a bit more about how the orthogonal behavioral task in Study 1 confirmed that participants were fully engaged in the FPVS presentation (p 7)?

10. When the Study 2 findings are first presented, it would be helpful to bring in some of the detail presented later on in the methods. Namely, I think it would be helpful to state that local culture members are the ones who determined which behaviors were relevant for each emotion category (i.e., that the 'contemporary culture' referenced on p 8 reflects Chinese culture). Relatedly, it could be explicitly stated that children's conceptual knowledge was not being assessed for how well it recapitulated cultural norms (i.e., how 'accurate' or 'aligned' it was), but simply how distinct the patterns of endorsement were across the various ratings.

11. Can the authors briefly interpret their Study 3 behavioral findings (children improved in their ability to distinguish between negative emotions over time, p 10) in light of their Study 2 findings (negative emotions came to be viewed as more similar with age)?

12. The matching task is the only one that is referenced with an acronym (MT) – this seems unnecessary; the task name is just as long as the sorting task.

13. I find the abbreviations of the emotion pairs, used in Figures 2 and 3, difficult to integrate. Is there a way to use the entire words, or shortened versions (e.g. hap-ang)?

References
Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. Psychological Bulletin, 143(3), 263–292. https://doi.org/10.1037/bul0000089
Boutonnet, B., Dering, B., Viñas-Guasch, N., & Thierry, G. (2013). Seeing objects through the language glass. Journal of Cognitive Neuroscience, 25(10), 1702–1710. https://doi.org/10.1162/jocn_a_00415
Gao, C., Oh, S., Yang, X., Stanley, J. M., & Shinkareva, S. V. (2024). Neural representations of emotions in visual, auditory, and modality-independent regions reflect idiosyncratic conceptual knowledge. Human Brain Mapping, 45(14), e70040. https://doi.org/10.1002/hbm.70040
Hoemann, K., Xu, F., & Barrett, L. F. (2019). Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. Developmental Psychology, 55(9), 1830–1849. https://doi.org/10.1037/dev0000686.
Jack, R. E., Sun, W., Delis, I., Garrod, O. G. B., & Schyns, P. G. (2016). Four not six: Revealing culturally common facial expressions of emotion. Journal of Experimental Psychology: General, 145(6), 708–730. https://doi.org/10.1037/xge0000162
Keller, H., & Otto, H. (2009). The cultural socialization of emotion regulation during infancy. Journal of Cross-Cultural Psychology, 40(6), 996–1011. https://doi.org/10.1177/0022022109348576
Nook, E. C., Sasse, S. F., Lambert, H. K., McLaughlin, K. A., & Somerville, L. H. (2017). Increasing verbal knowledge mediates development of multidimensional emotion representations. Nature Human Behaviour, 1(12), 881–889. https://doi.org/10.1038/s41562-017-0238-7
Nook, E. C., Sasse, S. F., Lambert, H. K., McLaughlin, K. A., & Somerville, L. H. (2018). The nonlinear development of emotion differentiation: Granular emotional experience is low in adolescence. Psychological Science, 29(8), 1346–1357. https://doi.org/10.1177/0956797618773357
Ogren, M., & Sandhofer, C. M. (2022). Emotion words link faces to emotional scenarios in early childhood. Emotion, 22, 167–178. https://doi.org/10.1037/emo0001063

Price, G. F., Ogren, M., & Sandhofer, C. M. (2022). Sorting out emotions: How labels influence emotion categorization. Developmental Psychology, 58, 1665–1675. https://doi.org/10.1037/dev0001391
Vigliocco, G., Kousta, S.-T., Della Rosa, P. A., Vinson, D. P., Tettamanti, M., Devlin, J. T., & Cappa, S. F. (2014). The neural representation of abstract words: The role of emotion. Cerebral Cortex, 24(7), 1767–1777. https://doi.org/10.1093/cercor/bht025
Wang, Q. (2001). '"Did you have fun?"' American and Chinese mother–child conversations about shared emotional experiences. Cognitive Development, 23.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
This is an interesting, original, elegant and well-written manuscript. It aims to examine to what extent behavioural judgment about facial expressions (i.e. behavioural facial expression discrimination) is determined by neural/perceptual sensitivity for facial features or by conceptual knowledge about emotions, and whether these associations change throughout development, covering samples of 5-to-10-year-old children and adults. The results suggest that the impact of perceptual abilities diminishes, while the impact of conceptual knowledge increases throughout development.

I did enjoy reading the manuscript, but I have the following concerns, which should be addressed before it can be considered for publication.

- I am convinced about the robustness and reliability of the frequency-tagging EEG approach to pinpoint perceptual discrimination abilities, even in very young children. Yet, I am uncertain about the reliability and validity of the conceptual knowledge rating task in child populations, especially in pre-schoolers who may not even be reading. It seems me that this task does not only taps into language (reading?) skills but also requires quite advanced meta-cognitive and abstract thinking and generalizing abilities, and even some intuitive sensing of the mental number line. Evidently, all these abilities are still very immature and fragile in pre-schoolers or in young kids entering primary school, which may result in highly unreliable and noisy data in this subpopulation (which may present as fuzzier less nuanced emotional representations). Accordingly, I wonder whether the increasing impact with age of conceptual emotional knowledge (as measured by this rating task) on behavioural facial emotion discrimination may not merely be the result of the increased reliability of the measures?
- On a related note, there is quite some evidence that performance on classical facial emotion discrimination tasks is correlated with verbal intelligence. To what extent may this conceptual knowledge rating task also tap into verbal IQ, and may the mutual relationship be driven by underlying individual differences in verbal IQ?
- In the Results section, I would recommend the authors to provide more complete statistical information. This implies a more detailed description of the models, reporting statistics for all the assessed effects (not only the significant results), applying formal post-hoc testing incorporating multiple-comparison corrections, etc. I noticed that some of the additional information could be found in the Suppl Information, but even then, it is not complete and sometimes even more confusing because of its fragmentary nature. The current statistical information is insufficient to fully evaluate the findings.
- In contrast with almost all other FPVS frequency-tagging EEG studies, the authors did not use the classical baseline-subtracted amplitude as key outcome measure, but they calculated signal-to-noise ratios (SNR). The reasoning is that the baseline-subtracted amplitudes revealed an age-dependent decline in power (possibly due to differences in scalp thickness), and therefore this index may not allow to study developmental changes in neural sensitivity for facial emotion discrimination. Yet, I may recall from the initial papers of the team of Bruno Rossion that these SNRs cannot simply be averaged or summed over the different harmonics. Therefore, I would urge the authors to verify the validity of this approach, in particular because of the very low correlations that they report between baseline-subtracted amplitudes and SNR measures (r=0.222).
- Also, for the youngest age groups, the interpretation differs significantly whether one examines SNRs (Fig. 2C) or baseline-subtracted amplitudes (SI Fig. 2). According to the first analysis approach, pre-schoolers can already differentiate among all the negative emotions, whereas according to the second analysis approach, they are not at all able to differentiate among these three negative emotion facial expressions.
- How were these SNR data integrated across multiple harmonics? Why are SNR data averaged across different harmonics for the 5-to-8-year-olds versus the 9-to-10-year olds and adults? What impact may this have on potential developmental effects? In addition to the group-averaged data, it would be informative to know what proportion of the participant group showed significant oddball discrimination responses (hence at the individual subject level).
- It is suboptimal that the FPVS segments comprised 20 s in the youngest age group and 30 s in all the older age groups, because less data generally results in less sensitive neural discrimination measures. The authors dealt with this issue by also recalculating the neural discrimination responses in the older age group while only using the first 20s segments, and they provide this as a supplementary analysis. In my opinion, this should be the main analysis which should be reported in the main text of the article.
- What is the rationale of additionally controlling for FACS defined facial stimulus differences? Isn't this a stricter control for the perceptual discrimination predictor as for the conceptual knowledge predictor?
- In the adult sample, I think it is striking that neural sensitivity for facial expression discrimination does no longer relate to behavioural facial expression discrimination, whereas the conceptual knowledge measure does (especially since conceptual knowledge is not really involved in the behavioural tasks, certainly not in the matching task). How can you explain this? Can it be because all commonalities among the neural and behavioural facial expression discrimination tasks

have been explained away by entering this FACS facial stimulus features?
- Regarding the conceptual knowledge measures: With increasing age, the various negative emotions become more intertwined, which the authors consider a positive maturation effect. Is this the case, as it may point towards less distinct emotional representations?

Minor issues:
- The manuscript is well-written, but I did spot some minor typos which should be corrected.
- It seems that all age groups were matched for gender, but it may be relevant to explicitly specify this (given that also gender differences in emotional processing have been demonstrated).
- There is no need to abbreviate the matching task and refer to it as MT. Please, refer to it in full.
- L61-71: "Therefore, children's developing ability …" ==> complex phrasing
- L108: add "… automatic structural encoding AND DISCRIMINATION of the stimuli …"
- L136: I am not fully convinced that the two tasks of Study III really assess "children's understanding of emotions". In my opinion, they primarily assess behavioural facial expression identification/discrimination/matching across two different operationalizations.
- L136 and following: Here, also, it would be helpful for the reader if you already shortly describe the two tasks that are used as the dependent variables of the regression models.
- L191-192: "Indicating that their responses at the odd frequencies were stronger than at the base frequencies." ==> This statement is incorrect. Responses at the odd frequencies are larger than the average response of the 16 surrounding bins, but not necessarily larger than the base rate 6Hz responses.
- L195-205: Please, report full statistics of the models, including the specific post-hoc (corrected?) comparisons.
- L209-220 (contrasting upright vs inverted faces): Please, provide full statistics for this analysis. E.g. was there a group x condition (upright vs inverted) interaction?
- L261 "maximum deviation parameter": Has this parameter been validated already? For instance, to what extent is it also determined by (more immature) motor coordination, which probably also correlates with age?
- L274 "the reciprocal of the SNR": How was this crucial measure exactly calculated? Across what channels / ROIs / harmonics …?
- Facial Action Coding System: I am a bit surprised that the authors still used the manual coding, whereas nowadays there exist well-validated computer vision software approaches that can perform the job (e.g. FaceReader, OpenFace, …).
- L309-315: Please, provide full statistical information for the models on the adults, including statistical information on the perceptual discrimination models. Can the authors speculate why neural discrimination scores no longer predict behavioural expression discrimination in adults (e.g. may neural scores be at ceiling and lack variability to predict behaviour?)?
- L493-499: The authors apply a new pre-processing method where they remove 10s epochs of data that do not show good enough quality. While the approach may be valuable, it may result in less data (and thus less robust responses) in the younger age groups. Can the authors present an overview of the number of data segments that were ultimately included for each of the age groups? See in this respect also my previous comment that the main analysis should be the one that is calculated on the 20s segments for all participants.
- L544: "… Pearson correlations between children's and adult's ratings …" ==> Where are these correlations reported?
- L565-566: Verify the order of the examples in this sentence. It does not seem to fit.
- Fig 4B, 4D: Better to change low level, middle level and high level of age to younger, middle and older age group, respectively.
- SI L124, paragraph 3.3: What do these KDEF stimuli come to do here?
- SI Fig. 5: Provide more information on the exact analysis approach. E.g., how were these correlations integrated across multiple emotion pairs?
- SI Fig. 6: Provide the units of the Y-axis for Fig A and B.
- SI Fig.12: Please, also display the results of the non-significant prediction of behavioural emotion judgements by neural perceptual discrimination.

I thank the authors and editors for giving me the opportunity to read and evaluate this manuscript.

Kind regards,
Bart Boets

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have produced a substantial revision that addresses most of my original concerns. I have just a few small points remaining.
- The abstract makes it seem like there are several studies each with 120-140 Ps… could you rephrase to clarify that these tasks were all implemented on one sample?
- I feel like my response to my prior point 5 is still not resolved… Do EEG data suggest that young kids differentiating negative emotion expressions from each other or not? The main text claims "the ability to discriminate facial configurations

has already emerged by 5 to 6 years of age," but I don't think either the main text analyses nor Supplemental Tables 4A-4G present the specific result I'm curious about (do young kids distinguish *negative* expressions). This feels like an important test to run and an important result to be clear about, as it weighs on lots of debate in the field about child representations of emotion being primarily about valence vs. specific emotions.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
My thanks to the authors for their attentive responses to my comments – many of which have been fully addressed. There are a few that I think could use some more clarification or elaboration.

In my previous point 1c, I questioned whether valence-based distinctions could be conceptual as well as perceptual. In their response, the authors note that their definition of valence highlights that it "is not merely a perceptual process but represents an integration of conceptual understanding that develops through children's embodied and social experiences." This is lovely phrasing – can it be incorporated into the text? There are also references to valence as a 'basic emotional cue' or 'sensory information' in the discussion (pp 15-16) that feel contradictory to an idea of it as integrated conceptual understanding. Can the authors further harmonize their thinking?

As part of my previous point 2, I wondered if the authors could state why they included only one positive emotion. I did not see a response to this point and think it would be useful to specifically point out this aspect when mentioning future directions (p 18).

In response to my previous point 10, the authors helpfully clarified that, in Study 2, children's conceptual knowledge was not being assessed for how well it aligned with cultural norms. However, there is a comment in the concluding paragraph about "the importance of exposure to cultural norms related to emotion" that now seems out of place.

More generally, I have the impression that the proposed developmental pathway for emotion understanding, discussed on page 15 of the discussion, is less clear in the revised manuscript than in the original. This has partially to do, I imagine, with the revisions in response to my point 11 (regarding comparison of Study 2 and 3 results). There is, for me, an uneasy tension between children's concepts becoming more complex (i.e., overlapping) in their features and simultaneously more differentiated in their deployment. Can the authors further specify their thinking here (e.g., via examples)?

I was asked to review the authors' responses and revisions to the comments from Reviewer 3. I confirm that the present comments can be shared with the authors.

Overall, I find that the authors responded exceedingly thoroughly and thoughtfully to the Reviewer's comments. They undertook several additional rounds of analyses and replaced/expanded many of the original results. They also took care to anchor their methodological decisions in the literature, especially regarding FPVS. I found their arguments compelling but was also happy to see them point to open questions that can be addressed by future research.

I have three small comments in response:
- In response to Reviewer 3 comment 1, the authors marshal evidence to suggest that the findings from the conceptual knowledge task were not impacted by younger children's cognitive abilities – even though they acknowledge that older children are generally more advanced in this domain. I didn't see this thinking reflected in the manuscript (e.g., as a response to potential alternative explanation in the discussion). Is there room to add it?

- Similarly, in response to Reviewer 3 comment 20, the authors clarify that "MD primarily reflects the trajectory of mouse movement as it relates to competing decision alternatives, rather than raw motor accuracy or speed". Can this be added to the methods section where this measure is introduced?

- Finally, in response to Reviewer 3 comment 10, the authors state that older children's integration across negative emotion concepts "does not imply reduced differentiation but rather a more interconnected conceptual framework". In line with my comments on this revision, I am wondering if this language – and some examples - can be added to the discussion to help clarify this point. It is an important one.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The reviewers have addressed all my remaining concerns.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I was asked to review the authors' responses and revisions to my own comments (Reviewer 2) as well as those from Reviewer 3. I confirm that the present comments can be shared with the authors.

My thanks again to the authors for their conscientious and articulate revisions to the manuscript. I have no further comments and look forward to seeing this paper in print.

(Remarks on code availability)
I have previous provided comments on the code.

RESPONSE TO REVIEWERS' COMMENTS

Reviewer #1 (Remarks to the Author):

This is an interesting paper that has several strengths. First, it uses a battery of tasks to probe several components of emotion development. Second, it tackles these questions in China, adding to the diversity of our typical participant pools. Third, it uses both brain and behavioral data.

Response: We appreciate the positive feedback on our paper.

Major concerns

I had the most questions about the conceptual knowledge task.
1. In the main text, it makes it seem like the ratings were between two emotion words (e.g., anger and sadness), but in the methods the task is described as rating similarity between emotion words and other relevant terms (e.g., anger and crying). Which is it, as these seem like very important differences.

Response: We clarified this issue in the revised manuscript. Participants rated the association of each term (e.g., "crying") with the four emotion categories (happiness, anger, sadness, and fear). These ratings provided insight into how strongly each term was linked to each emotion category. To measure the conceptual similarity among the categories (e.g., anger and sadness), we calculated Pearson correlation coefficients between the ratings of each emotion pair using the 20 terms for each participant. Please see this revised text on Page 9, Lines 256-259 and Page 22, Lines 692-696 (in the changes markup version of the MS).

2. The task seems to be analyzed using a series of correlation coefficients which are then averaged together. It seems like the more conventional way to analyze a task like this (Barrett, 2004, JPSP; Nook et al., 2014, NHB; Suvak et al., 2011, Abnormal). It's strange that negative emotions would be seen as more similar as age increases, so I'm curious if analyses using an MDS approach provides a different set of results or the same.

Response: Following the reviewer's suggestion we additionally applied multidimensional scaling (MDS) to examine pairwise Euclidean distances between emotions. Overall, using an MDS approach yielded findings nearly identical to those we reported in the manuscript. Specifically, we computed dissimilarity matrices for each participant with Euclidean distances and reduced them to a two-dimensional representation (mean stress < 0.1). A repeated measures ANOVA using this MDS measure replicated the main results from the analysis with correlation coefficients, showing a significant interaction between age and emotion pairs ($F(10.413, 621.300) = 2.223$, $p = .014$, $\eta_p^2 = 0.036$). Post-hoc analyses also revealed that the conceptual dissimilarities between negative emotions decreased with age. The only difference was that the post-hoc analyses did not replicate the age-related decrease in similarity between positive and negative emotions found with correlation coefficients. The analysis with Euclidean distances also replicated the main effect of emotion pair, $F(3.471, 621.300) = 358.640$, $p < .001$, $\eta_p^2 = 0.667$, with happiness consistently perceived as the most distinct from negative emotion categories. We have added these MDS results as SI Figure 6 and Tables 7A, 7B and 7C to Page 4-5, Lines 95-115, Page 18, and Page 40-42 in the SI document.

We chose the more conventional measure of correlation coefficients because it is more robust against response biases, such as the consistent use of extreme ends of the rating scale, compared to approaches like Euclidean distances (Hu et al., 2022, NeuroImage). Additionally, correlation analysis provides information about positive vs. negative relationships and better reflects intra-subject emotion relationships, while the measure of Euclidean distances does not. This additional information could provide deeper insights into the development of conceptual knowledge over childhood.

3. This result of increased similarity across age for negative emotions also seems to emerge only in a correlation where age is rounded to the nearest integer (SI Fig 5). Can you use an unrounded age estimate for continuous age analyses? (e.g., if someone is 5 years, 6 months, and 15 days old, they're 5.55 years old).

Response: Yes, we have reanalyzed the data according to this comment using unrounded continuous age values and have reported the results in the revised manuscript. The analysis resulted in consistent findings with our previous reported data: the conceptual similarity between happiness and negative emotions did not exhibit a statistically significant relationship with age, whereas the similarity among negative emotions continued to show a significant increase with age. To keep consistency, we have also applied this continuous age calculation across all analyses involving age. Please see Page 3-4, Lines 83-94 in the SI document.

4. In several places, this task is described as assessing "norms" about emotions. Although I'm completely on board with the assumption that kids' conceptual knowledge of emotions is learned from the social milieux, I don't think a given individual's semantic understanding of emotions is a direct measure of a group's norms. There can be deviance between what one person thinks and what a group tends to think, no? I'd recommend clarifying or removing these points.

Response: Thank you for this thoughtful comment. We agree that an individual's conceptual knowledge of emotions cannot be equated with group norms. In response, we have revised the references to "norms" and instead clarified that our task measures how children's semantic understanding of emotions may be influenced by their cultural and social environments, including exposure to stereotypes, rather than directly assessing collective and broader societal norms. These edits appear on Page 9, Lines 243-244, Page 14, Line 411, and Page 16, Lines 453.

5. I was also left unsure of whether young children did or did not seem to perceptually discriminate negative emotions according to EEG data. The analyses clearly show that pos/neg faces are discriminated and that neg discrimination increases with age. But are (for example) oscillating sad and angry faces producing an oddball effect in 5yos? . . .

Response: We concur with the reviewer on this point and revised the text to reflect that perceptual discrimination emerges by age 5 to 6 years rather than characterizing it as firmly established. Please see Page 8, Lines 222-223.

6. How many times did the following occur: "The experimenter would skip a particular item if the child still found an item incomprehensible after the explanation" across each age? It's important to note exclusion rates.

Response: That situation was rare, occurring in less than 6% of cases, even in the youngest group. We have now included the exclusion rates in the revised manuscript on Page 22, Lines 635-638.

Smaller concerns

7. The abstract is quite jargony. It's not clear how discrimination and conceptual knowledge are operationalized, making it difficult for readers to get a sense of what was done and what was found.

Response: We have revised the abstract to clarify how discrimination and conceptual knowledge are operationalized.

8. Terminology was difficult throughout, actually. MT, sorting, conceptual knowledge, etc. are fairly abstract and unintuitive terms. . ., I'd recommend clearly laying out each construct you're interested in and naming how you measure each. . .

Response: In the revised manuscript we reduce use of acronyms and use the full names of each task. In addition, we clarify how each construct is measured and explicitly link the tasks to the constructs they assess. Please see Page 2, Lines 35-41, Page 5, Lines 128-136 and Page 11, Lines 325-326.

9. If these "studies" are all on the same children, are they different studies? It seems to me that this is one study with several tasks.

Response: We greatly appreciate this question. The authors have carefully considered and discussed replacing all instances of "studies" with "tasks." Upon reflection, we do think it is easiest for readers to segment the paper based upon designation of different studies (each encompassing one or more tasks). This is because, as traditionally defined, each study explores a distinct hypothesis and we believe the current organization will best help readers distill and synthesize the data and argument of the paper.

10. The intro follows a strange structure in that the study is introduced several times. Typically, the intro reviews each section of background literature, culminating in an open research question. This is then used to introduce the study just once at the end as a way to answer this question.

Response: We restructured the introduction to align with this suggested format.

11. In intro, sentence 2 is written in a way that makes it seem like it's in tension with sentences 3-4 but aren't they compatible?

Response: We appreciate the opportunity to clarify this point. We aimed to contrast two perspectives in the literature. Sentence 2 reflects the traditional view that emotional development progresses through increasing sensitivity to perceptual cues (such as facial configurations), supporting the idea of universal facial "expressions" across cultures. In contrast, the subsequent sentences suggest that perceptual cues alone are insufficient. This latter perspective posits that conceptual knowledge, shaped by cultural and linguistic input, plays a critical and complementary role. We have clarified it in the revised manuscript.

12. In this sentence, the terminology is not clear "due to the processing of structural changes between facial configurations rather than featural changes". What exactly is meant by facial configurations vs featural changes? The inverted oddball control is actually extremely important for the logic of this study.

Response: Thank you for highlighting this important point. We have clarified the distinction between "facial configurations" and "featural changes" in the manuscript. Specifically, facial features refer to individual elements such as the shape of the eyes or mouth and the presence of wrinkles, while facial configurations denote the spatial arrangement and relative positioning of these features (e.g., the distance between the eyes, nose, and mouth). Inverted images disrupt the facial configural information but preserving featural details. Thus, results from the inverted oddball control experiment confirm that the oddball responses to upright faces reflect the perceptual processing of structural changes between facial configurations rather than just isolated featural differences. Please see Page 8, Lines 225-233.

13. Beware causal language throughout, as the study is correlational (e.g., "conceptual knowledge significantly influenced the results" in SI)

Response: We have carefully revised the manuscript and supplementary materials accordingly to avoid causal

implications.

14. Line 77 – are they really "experiments"?

Response: We have replaced all instances of "experiments" with "tasks".

15. Abstract N should be final usable sample, not the sample you collected (30 excluded)

Response: We have updated the abstract to reflect the final usable sample sizes for each task after exclusions.

Reviewer #2 (Remarks to the Author):

I found this an elegant set of studies that nicely builds on constructionist theories and recent empirical work (namely Brooks & Freeman, 2018) to illustrate how domain-general processes shape emotional development in different but intersecting ways, as children move through different ways of interfacing with the world. 'Emotional' input (that is, inputs that scaffold interactions with others) is at first largely perceptual (e.g., based on facial movements), and becomes more conceptual (e.g., anchored in language) over time. With this in mind, I think there are a few ways the authors could make the value of their contribution more evident and accessible to a broader set of readers.

Response: Thank you for your thoughtful and encouraging feedback.

1. More definition and explanation of the 'perceptual' vs 'conceptual' distinction would help clarify the nature of the processes at play. . .

Response: Thank you for raising this important point. We agree that perceptual processes can indeed incorporate top-down influences of conceptual knowledge, as demonstrated in adults – sometimes even without conscious awareness (e.g., Boutonnet et al., 2013, JoCN; Brooks et al., 2019, PNAS). However, the extent to which conceptual knowledge of emotion modulates perceptual discrimination in children is less established, at least within the age range tested in our study. Based upon the Reviewer's comment, we explored the relationship between perceptual discrimination and conceptual knowledge but found no significant association. While we acknowledge that null results should not be over-interpreted, as the editor pointed out, this suggests that the top-down influence of conceptual knowledge on perceptual processes may be limited during this developmental period. We have clarified this point in the revised manuscript on Page 10, Lines 277-282 (in the changes markup version of the MS).

b. In Study 1, the authors confirm that faces were being processed configurally, as inverted faces elicited lower oddball effects. How, if at all, does the distinction between featural and configural processing map onto the distinction between perceptual and conceptual discrimination? Does configural processing imply the 'concept' of a face?

Response: We consider both featural processing and configural processing as important aspects of perceptual discrimination. The reduced oddball responses to inverted faces observed in our study highlight the critical role of configural processing during perceptual discrimination, as noted in our response to Reviewer 1's comment #12. However, while face inversion disrupts configural processing, it should not prevent individuals from perceiving and conceptualizing the inverted image as a face. For these reasons, we do not think the distinction between featural and configural processing directly corresponds to the distinction between perceptual and conceptual discrimination. We have clarified the interpretation of the inverted control experiment based on your comment and Reviewer 1's related feedback. Please see our changes above.

c. There are also a few points where it is implied that broad, valence-based distinctions between facial configurations are 'perceptual' distinctions – or, at least, that they are not yet sophisticated enough to count as 'conceptual'. This implication does not align with the thinking in many of the constructionist papers cited here. Instead, valence is conceived of a broad conceptual dimension that children may leverage to distinguish between faces, based presumably on their embodied experiences of encountering these faces in the interactions with caretakers and others around them.

Response: We appreciate the Reviewer's insightful comment. We agree that valence-based distinctions could reflect children's conceptual knowledge and be shaped by experiences with those facial configurations. We recognize that some ambiguity may have arisen from the phrasing in our introduction and discussion. Thus, we have edited the text where we explicitly defined valence as the broad judgment of emotional configurations as "feels good or bad." This foundational distinction serves as a basis for children's emerging ability to differentiate between more specific emotions. Importantly, this definition highlights that valence is not merely a perceptual process but represents an integration of conceptual understanding that develops through children's embodied and social experiences.

2. Can the authors say more about why they selected the four emotion categories that they did (anger, fear, happiness, and sadness), what implications this has for the present work? I am curious, for example, why only one positive emotion category was included, and why categories that (at least in some cultural groups) are known to develop later on (e.g., disgust, shame) were not. Relatedly, do these four categories map onto this identified by Jack and colleagues (2016)?

Response: We selected the four emotion categories based on developmental and methodological considerations. These categories have the most current empirical data showing advancements in differentiation and understanding during the preschool to middle childhood years (Widen, 2013). Methodologically, limiting the number of categories helped sustain young children's attention and ensured robust data collection, particularly for the EEG FPVS task. Including more emotions risked increasing cognitive load and reducing engagement. Our selected categories differ slightly from Jack and colleagues (2016). While Jack and colleagues identified broader latent dimensions that include surprise-like expressions, we did not include this category in the current study and acknowledged this as a direction for future research to expand our understanding of emotion development across cultures. Please see Page 18, Line 532-536.

3. I wasn't quite sure how to interpret the results of Study 1. It is stated both that "some emotion categories were better discriminated with increasing age" (p 7) and that "there were no significant differences among the three child groups" (p 8). There is also a reference to behavioral findings being reflected in EEG activity (p 7); what are the behavioral findings being referenced – are they from this study?

Response: We clarified these points and also reanalyzed the data based on suggestions from Reviewer 3's comment # 6 and 7. Specifically, we selected the same harmonics across all conditions and age groups and used the same length of trials per age group. This analysis revealed an interaction between ROI and age, indicating that oddball responses increase with age in the middle and right temporal-occipital regions. There was also a main effect of emotion. However, the interaction between emotion and age was not significant. These findings suggest the ability to discriminate facial configurations has already emerged by 5 to 6 years of age. Thus there is a general enhancement in the perceptual processing of facial configurations with age, but this increase is not specific to particular emotion categories.

Regarding the statement "behavioral findings being reflected in EEG activity," we acknowledge that this was an error in phrasing. The results discussed in this section are based solely on EEG findings and do not refer to any behavioral results. We have revised the text to accurately reflect this and ensure clarity in the presentation of our findings. These revisions may be found on Page 8, Lines 212-224.

4. The interpretation of Study 3 (and the paper overall) is that perceptual discrimination influences emotion judgment earlier on in childhood, but that conceptual knowledge comes online to dominate later in childhood. The actual results, however, are more complex than this. In the sorting task, there is no interaction between conceptual knowledge and age (only a positive main effect). In the matching task, there is no interaction between perceptual discrimination and age (only a positive main effect). So it seems an alternative explanation is that these tasks differentially tap into the proposed underlying mechanisms – that there is no cohesive intersection of the mechanisms that extends across all contexts for emotion judgment. I'd like to see the authors engage with these nuances in their discussion.

Response: We appreciate this insightful comment. We have now revised our summary of the findings from two tasks and elaborated on this point in the discussion to address the complexity of the findings according to your comment on Page 16, Lines 467-473.

5. This work inspires a lot of connections with the existing literature, both developmental and otherwise. To the extent possible within this submission format, the authors could consider integrating the following references:
a. "cultural frameworks for displaying emotions" (Keller & Otto, 2009; Wang, 2001) b. "linguistic input about emotions" (Borghi et al., 2017; Hoemann et al., 2019; Nook et al., 2017; Vigliocco et al., 2014) c. "evidence that conceptual knowledge guides children's reasoning about emotion" (Ogren & Sandhofer, 2022; Price et al., 2022) d. "older children will show more distinct neural responses between configurations" (Gao et al., 2024) e. "[children] conceptualize various negative emotions as more similar to each other with age" (Nook et al., 2018)

Response: Thank you for your positive feedback and providing these references, which we found very helpful in strengthening our manuscript. We have reviewed each of them carefully and incorporated those that align closely with our study's focus into the revised text.

Minor miscellaneous points:

6. Please briefly define "valence-based judgement" (p 3).

Response: We have added a brief definition of "valence-based judgment" in the manuscript, (i.e., feels good or bad; Widen, 2013).

*7. Given what we know about emotional development across the lifespan, I might say that children development progresses to *more* nuanced negative emotion differentiation (p 3) – in other words, there are still levels of nuance yet to come!*

Response: We have revised the manuscript to better highlight this developmental continuum as the Reviewer suggested, i.e., "more nuanced negative emotion differentiation".

8. A typo in line 77 states that 4 experiments were conducted (instead of 3).

Response: We have corrected it accordingly.

9. Can the authors say just a bit more about how the orthogonal behavioral task in Study 1 confirmed that participants were fully engaged in the FPVS presentation (p 7)?

Response: We have added more details in the manuscript on Page 7, Lines 193-197 to clarify how the orthogonal

behavioral task helped participants' engagement in the FPVS presentation. Additionally, we have included an example of the black-and-white image in Figure 1A to provide a visual illustration.

10. When the Study 2 findings are first presented, it would be helpful to bring in some of the detail presented later on in the methods. Namely, I think it would be helpful to state that local culture members are the ones who determined which behaviors were relevant for each emotion category (i.e., that the 'contemporary culture' referenced on p 8 reflects Chinese culture). Relatedly, it could be explicitly stated that children's conceptual knowledge was not being assessed for how well it recapitulated cultural norms (i.e., how 'accurate' or 'aligned' it was), but simply how distinct the patterns of endorsement were across the various ratings.

Response: We have revised the manuscript to include this information when the findings are first presented. Additionally, we explicitly state that the conceptual knowledge task was not designed to assess how well children's ratings aligned with cultural norms, but rather to evaluate the distinctiveness of their conceptual associations across different emotions. This clarification aligns with our response to Reviewer 1's comment #4 regarding a similar point.

11. Can the authors briefly interpret their Study 3 behavioral findings (children improved in their ability to distinguish between negative emotions over time, p 10) in light of their Study 2 findings (negative emotions came to be viewed as more similar with age)?

Response: While the findings from Studies II and III may seem contradictory, we interpret them as reflecting different facets of emotional development. Study II shows that with age, children perceive negative emotions as more conceptually similar, suggesting a more integrated and organized understanding of negative emotional experiences. This likely reflects higher-order categorization, where children recognize shared features and contexts across negative emotions, leading to stronger clustering without reducing differentiation. In contrast, the improved behavioral differentiation in Study III's indicates that, despite perceiving negative emotions as conceptually more related, older children become better at identifying subtle distinctions between them in context. This suggests that as children's emotional knowledge becomes more interconnected, they gain greater flexibility and precision in distinguishing emotions in real-world contexts. This conceptual refinement and a more interconnected emotional framework actually enhance, rather than hinder, fine-grained discrimination in behavioral contexts. We have elaborated on this interpretation in the Discussion.

12. The matching task is the only one that is referenced with an acronym (MT) – this seems unnecessary; the task name is just as long as the sorting task.

Response: We have revised the manuscript to use the full term, "the matching task," throughout, aligning it with the naming convention for the sorting task.

13. I find the abbreviations of the emotion pairs, used in Figures 2 and 3, difficult to integrate. Is there a way to use the entire words, or shortened versions (e.g. hap-ang)?

Response: To address this, we have replaced all abbreviated emotion pairs in Figures 2 and 3 with their full names (e.g., "happiness-anger") as you suggested.

References

Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychological Bulletin, 143*(3), 263–292. https://doi.org/10.1037/bul0000089

Boutonnet, B., Dering, B., Viñas-Guasch, N., & Thierry, G. (2013). Seeing objects through the language glass. *Journal of*

Cognitive Neuroscience, 25(10), 1702–1710. https://doi.org/10.1162/jocn_a_00415
Gao, C., Oh, S., Yang, X., Stanley, J. M., & Shinkareva, S. V. (2024). Neural representations of emotions in visual, auditory, and modality-independent regions reflect idiosyncratic conceptual knowledge. *Human Brain Mapping, 45*(14), e70040. https://doi.org/10.1002/hbm.70040
Hoemann, K., Xu, F., & Barrett, L. F. (2019). Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Developmental Psychology, 55*(9), 1830–1849. https://doi.org/10.1037/dev0000686.
Jack, R. E., Sun, W., Delis, I., Garrod, O. G. B., & Schyns, P. G. (2016). Four not six: Revealing culturally common facial expressions of emotion. *Journal of Experimental Psychology: General, 145*(6), 708–730. https://doi.org/10.1037/xge0000162
Keller, H., & Otto, H. (2009). The cultural socialization of emotion regulation during infancy. *Journal of Cross-Cultural Psychology, 40*(6), 996–1011. https://doi.org/10.1177/0022022109348576
Nook, E. C., Sasse, S. F., Lambert, H. K., McLaughlin, K. A., & Somerville, L. H. (2017). Increasing verbal knowledge mediates development of multidimensional emotion representations. *Nature Human Behaviour, 1*(12), 881–889. https://doi.org/10.1038/s41562-017-0238-7
Nook, E. C., Sasse, S. F., Lambert, H. K., McLaughlin, K. A., & Somerville, L. H. (2018). The nonlinear development of emotion differentiation: Granular emotional experience is low in adolescence. *Psychological Science, 29*(8), 1346–1357. https://doi.org/10.1177/0956797618773357
Ogren, M., & Sandhofer, C. M. (2022). Emotion words link faces to emotional scenarios in early childhood. *Emotion, 22*, 167–178. https://doi.org/10.1037/emo0001063
Price, G. F., Ogren, M., & Sandhofer, C. M. (2022). Sorting out emotions: How labels influence emotion categorization. *Developmental Psychology, 58*, 1665–1675. https://doi.org/10.1037/dev0001391
Vigliocco, G., Kousta, S.-T., Della Rosa, P. A., Vinson, D. P., Tettamanti, M., Devlin, J. T., & Cappa, S. F. (2014). The neural representation of abstract words: The role of emotion. *Cerebral Cortex, 24*(7), 1767–1777. https://doi.org/10.1093/cercor/bht025
Wang, Q. (2001). "'Did you have fun?"' American and Chinese mother–child conversations about shared emotional experiences. *Cognitive Development, 23*.

Reviewer #3 (Remarks to the Author):

This is an interesting, original, elegant and well-written manuscript. It aims to examine to what extent behavioural judgment about facial expressions (i.e. behavioural facial expression discrimination) is determined by neural/perceptual sensitivity for facial features or by conceptual knowledge about emotions, and whether these associations change throughout development, covering samples of 5-to-10-year-old children and adults. The results suggest that the impact of perceptual abilities diminishes, while the impact of conceptual knowledge increases throughout development.

Response: We appreciate the Reviewer's thoughtful and encouraging comments on our manuscript.

1. I am convinced about the robustness and reliability of the frequency-tagging EEG approach to pinpoint perceptual discrimination abilities, even in very young children. Yet, I am uncertain about the reliability and validity of the conceptual knowledge rating task in child populations, especially in pre-schoolers who may not even be reading. It seems me that this task does not only taps into language (reading?) skills but also requires quite advanced meta-cognitive and abstract thinking and generalizing abilities, and even some intuitive sensing of the mental number line. Evidently, all these abilities are still very immature and fragile in pre-schoolers or in young kids entering primary school, which may result in highly unreliable and noisy data in this subpopulation (which may present as fuzzier less nuanced emotional representations). Accordingly, I wonder whether the increasing impact with age of conceptual emotional knowledge (as measured by this rating task) on behavioural facial emotion discrimination may not merely

be the result of the increased reliability of the measures?

Response: The items used in the task were selected by the children's guardians from an initial list of 40 options, resulting in the top 20 words they believed their child could read and understand. Additionally, during the experiment, if children encountered words they could not read or understand, the experimenter provided an explanation. If the child still could not comprehend the word, it was skipped. To further address this issue, the revised manuscript includes data detailing the proportion of skipped items due to incomprehension. The number of skipped items did not exceed 6% in any age group, demonstrating that the task was well-suited for pre-school and primary school-aged children. Please also see our response to Reviewer 1's comment # 6 asking for the same information. Additionally, our analyses with correlation coefficients and an alternative approach using Euclidean distances (see response to Reviewer 1's comment #2) consistently showed that the 5-6 years group demonstrated a similar pattern as the older groups. Taken together, while we acknowledge that older children generally have better cognitive and abstract-thinking abilities, these differences did not appear to significantly affect the younger children's performance in the task.

2. On a related note, there is quite some evidence that performance on classical facial emotion discrimination tasks is correlated with verbal intelligence. To what extent may this conceptual knowledge rating task also tap into verbal IQ, and may the mutual relationship be driven by underlying individual differences in verbal IQ?

Response: Unfortunately, we do not have a direct measurement of children's verbal IQ. However, we used a parent-reported child word comprehension questionnaire (Baron-Cohen, 2010), which included emotion-oriented vocabulary items related to the emotion categories examined in this study. Data were collected from the guardians of 89 children who participated in the study. To address this question, we analyzed whether children with better emotional word comprehension showed greater similarity to adults in the conceptual understanding task. After controlling for age as a covariate, no significant correlations were found between parent-reported emotional vocabulary scores and children's self-rated conceptual similarity ratings (happiness: r = -0.055, p = .610, df = 85; fear: r = -0.023, p = .829, df = 85; anger: r = 0.004, p = .972, df = 85; sadness: r = -0.124, p = .251, df = 85). These results have been added to the SI. Thus, it seems plausible that the results of the conceptual understanding task are unrelated to children's general verbal IQ. However, we acknowledge that parent-reported word comprehension has limitations in validity and is not equivalent to verbal IQ.

3. In the Results section, I would recommend the authors to provide more complete statistical information. This implies a more detailed description of the models, reporting statistics for all the assessed effects (not only the significant results), applying formal post-hoc testing incorporating multiple-comparison corrections, etc. . . .

Response: We have incorporated the Reviewer's recommendation by adding more detailed statistical information for the results in the main text. However, given the large number of pairwise comparisons across six emotion pairs in the post-hoc analyses, including all these results in the main text would significantly impact the manuscript's readability. To address this, we have now included the complete results in the SI, where they are presented in a clear table format for ease of interpretation.

4. In contrast with almost all other FPVS frequency-tagging EEG studies, the authors did not use the classical baseline-subtracted amplitude as key outcome measure, but they calculated signal-to-noise ratios (SNR). The reasoning is that the baseline-subtracted amplitudes revealed an age-dependent decline in power (possibly due to differences in scalp thickness), and therefore this index may not allow to study developmental changes in neural sensitivity for facial emotion discrimination. Yet, I may recall from the initial papers of the team of Bruno Rossion that these SNRs cannot simply be averaged or summed over the different harmonics. Therefore, I would urge the authors to verify the validity of this approach, in particular because of the very low correlations that they report between

baseline-subtracted amplitudes and SNR measures (r=0.222).

Response: We carefully revisited the relevant literature on FPVS and the usage of SNR measures and found that averaging SNRs across significant harmonics used for statistical analysis is an accepted approach in adult FPVS studies, as this measure corresponds to the same underlying phenomenon with baseline-corrected amplitude (e.g., Appelbaum et al., 2006; Heinrich et al., 2009; Liu-Shuang, Norcia, & Rossion, 2014; Alp et al., 2017). To address the validity of this method, we have added these references to the manuscript to substantiate our approach and highlight its reliability. Regarding the low correlation between baseline-corrected amplitudes and signal-to-noise ratio measures (r = 0.222), our initial analysis pooled data across all age groups, which may have reduced the correlation due to differences in how these measures vary with age. Specifically, baseline-corrected amplitudes tend to decrease with age, whereas SNR measures do not show the same trend, potentially explaining the overall low correlation. To address this, we conducted separate correlation analyses for each age group. The results revealed much higher correlations between baseline-corrected amplitudes and SNR measures within each group: 5–6 years (r = 0.336, $p = 0.024$, $df = 45$), 7–8 years (r = 0.547, $p < 0.001$, $df = 37$), and 9–10 years (r = 0.748, $p < 0.001$, $df = 39$). These findings have been added to the SI document on Page 1, Lines 23-26. In sum, we concur with the Reviewer that baseline-corrected amplitude is often used in adult research. However, our literature review suggests that SNR is also a common approach. More importantly, baseline-corrected amplitude has a much larger negative impact on pediatric data compared to SNR, particularly when a wide age range is explored. The latter is why we have conducted a comprehensive set of analyses to compare and justify the use of the two measures within our dataset.

5. Also, for the youngest age groups, the interpretation differs significantly whether one examines SNRs (Fig. 2C) or baseline-subtracted amplitudes (SI Fig. 2). According to the first analysis approach, pre-schoolers can already differentiate among all the negative emotions, whereas according to the second analysis approach, they are not at all able to differentiate among these three negative emotion facial expressions.

Response: After adopting the Reviewer's suggestion in comment #6 to use a consistent approach to average data with the same harmonics across all age groups, we now have observed that the 5-6 age group differentiated among all the negative emotions based on both measures (baseline-corrected amplitude >0, SI Figure2). This result indicates that preschoolers can already distinguish between all negative emotions, aligning with the findings derived from the signal-to-noise ratio measure and reinforcing the robustness of our conclusions. We suspect the original difference observed in the youngest group between the two measures may also (partially) stem from the inherent sensitivity of baseline-corrected amplitude to (random) extreme values in pediatric data. A detailed interpretation of differences between these measures across all analyses is beyond the scope of the current study, though we have added a note in the discussion suggesting that future research could explore which method is more suitable for specific contexts and participant age groups. These revisions may be found on Page 17, Lines 482-484 (in the changes markup version of the MS).

6. How were these SNR data integrated across multiple harmonics? Why are SNR data averaged across different harmonics for the 5-to-8-year-olds versus the 9-to-10-year olds and adults? What impact may this have on potential developmental effects? In addition to the group-averaged data, it would be informative to know what proportion of the participant group showed significant oddball discrimination responses (hence at the individual subject level).

Response: To address these important questions, we have done the following. First, we have adopted a consistent approach by averaging the data for the same harmonics across all age groups. Only overlapped harmonics that were significant across all age groups (i.e., 3.6Hz and 4.8Hz) were included in the analysis. This unified method minimizes bias and ensures a fair and reliable comparison of neural responses across development. We replicated all the analyses using this approach and found that the results were almost identical with the previous analyses. Second, we conducted an analysis at the individual subject level to determine the proportion of participants within each age group who

exhibited significant oddball discrimination responses. These findings, now included in the SI, demonstrate that over two-thirds of participants in each age group showed significant responses, further supporting the robustness of group-level results. These data may be found in SI document on Page 1, Lines 10-13.

7. It is suboptimal that the FPVS segments comprised 20 s in the youngest age group and 30 s in all the older age groups, because less data generally results in less sensitive neural discrimination measures. The authors dealt with this issue by also recalculating the neural discrimination responses in the older age group while only using the first 20s segments, and they provide this as a supplementary analysis. In my opinion, this should be the main analysis which should be reported in the main text of the article.

Response: To fully address this issue and ensure consistency across age groups, we have implemented the Reviewer's recommendation by re-performing all of our analyses with only the first 20 seconds of FPVS data for all participants, regardless of age. This adjustment applies to all analyses involving FPVS data, including the RSA, and yielded very similar findings after implementation. Additionally, we have now removed the results using the full 30 seconds of FPVS data for participants aged 7–10 years and adults to the SI.

8. What is the rationale of additionally controlling for FACS defined facial stimulus differences? Isn't this a stricter control for the perceptual discrimination predictor as for the conceptual knowledge predictor?

Response: The primary rationale for controlling FACS-defined facial stimulus differences is to ensure that the observed effects in the emotion understanding task are not confounded by low-level physical similarities/dissimilarities between stimuli. By accounting for FACS-defined differences, we can rule out the possibility that participants rely on simple visual resemblances (e.g., shared facial muscle movements) to make their judgments, which is critical for isolating the contributions of perceptual discrimination and conceptual knowledge. This is a common procedure used by studies on facial emotion perception (Gagnon et al., 2014; Brooks & Freeman, 2018). Without this control, there would be a risk that the relationship between predictors (perceptual and conceptual) and emotion understanding could be artificially inflated by the shared physical properties of the stimuli, as the same set of images were used across tasks (i.e., the Tsinghua facial expression database).

9. In the adult sample, I think it is striking that neural sensitivity for facial expression discrimination does no longer relate to behavioural facial expression discrimination, whereas the conceptual knowledge measure does (especially since conceptual knowledge is not really involved in the behavioural tasks, certainly not in the matching task). How can you explain this? . . .

Response: We appreciate the Reviewer's thoughtful questions. As we noted in the previous response, the FACS-defined parameters primarily account for some low-level physical similarities or dissimilarities between stimuli, which are far removed from the perceptual processing of a face. However, to test the Reviewer's hypothesis, we conducted an additional analysis to determine whether the results remained consistent after removing FACS variables from the model. We observed that the main effect of perceptual discrimination remained statistically non-significant for the matching task ($B = 0.087$, SE = 0.024, $Z = 20.83$, $p = .140$). This suggests that adults may exhibit ceiling effects in their perceptual discrimination of facial configurations for emotions (as pointed out in comment #23 down below), with individual differences in the task being more likely driven by their conceptual knowledge. This aligns with our broader developmental findings, where the influence of perceptual discrimination decreases with age while the role of conceptual knowledge becomes increasingly significant. For further details, the analysis excluding FACS variables has been included in the SI.

Regarding the question about conceptual knowledge in the matching task: While it is likely that the task reflects some degree of perceptual similarities between different facial configurations, the task instructions explicitly required participants to match the pictures based on how people are feeling emotionally (i.e., their emotional state). This instruction inherently encourages participants to infer and match emotional meaning, rather than merely focusing on perceptual similarities in facial configurations. Additionally, our findings show that adults perform significantly better than children in the matching task. Since the age effect was not as pronounced in the FPVS task, the difference in matching performance suggests that adults are better at incorporating higher-level conceptual knowledge into their judgments. We have added further explanation to the results and discussion sections to clarify these points.

10. Regarding the conceptual knowledge measures: With increasing age, the various negative emotions become more intertwined, which the authors consider a positive maturation effect. Is this the case, as it may point towards less distinct emotional representations?

Response: This is an insightful question that aligns with Reviewer 1's comment #2 and Review 2's comment #11. As we addressed in previous comments, our findings suggest that older children develop the ability to recognize shared features and contexts across negative emotions, leading to more coherent clustering in conceptual space. Importantly, this does not imply reduced differentiation but rather a more interconnected conceptual framework. Meanwhile, the results from the sorting task suggest that while older children perceive negative emotions as conceptually more related, they also improve in their ability to identify subtle distinctions between them in context. As suggested by our RSA results, this improvement likely reflects the growing precision and flexibility with which older children navigate their more integrated emotional knowledge, which enables them to differentiate emotions more effectively in real-world or task-specific contexts. This explanation also aligns with align with prior research (Nook et al., 2018, *Psychological Science,* suggested by Reviewer 2), showing that emotion differentiation follows a nonlinear trajectory: increasing overlap in emotional experiences during childhood, followed by greater precision with age.
We have expanded on this explanation in the discussion. Please also refer to our responses to Reviewer 2's comment #11.

Minor issues:

11. It seems that all age groups were matched for gender, but it may be relevant to explicitly specify this (given that also gender differences in emotional processing have been demonstrated).

Response: We have added a statement in the manuscript to explicitly clarify that all age groups were matched for gender.

12. There is no need to abbreviate the matching task and refer to it as MT. Please, refer to it in full.

Response: We have now replaced every instance of "MT" with "the matching task" throughout the manuscript.

13. L61-71: "Therefore, children's developing ability …" ==> complex phrasing

Response: We have revised the phrasing of this sentence to make it clearer and more accessible.

14. L108: add "… automatic structural encoding AND DISCRIMINATION of the stimuli …"

Response: We have updated this sentence accordingly.

15. L136: I am not fully convinced that the two tasks of Study III really assess "children's understanding of emotions". In my opinion, they primarily assess behavioural facial expression identification/discrimination/matching across two different operationalizations.

Response: We agree that two tasks of Study III may involve elements of facial expression identification and discrimination; however, we do not view this as conflicting with how we have defined the focus of these two tasks. These identification/discrimination processes are integral to emotion understanding (Ruba & Pollak, 2020, *Annual. Rev. of Dev. Psych*). More broadly, the tasks are designed to capture interrelated mechanisms that contribute to children's developing comprehension of emotions. Specifically, the sorting task emphasizes conceptual knowledge and higher-order reasoning in emotion understanding by asking children to categorize facial configurations into scenarios using scripts and labels, while the matching task captures both perceptual and conceptual aspects of emotion understanding and comparisons. These tasks have been widely adopted in developmental research for these purposes (Bayet et al., 2018; Brooks & Freeman, 2018; Rodger et al., 2015). Meanwhile, we acknowledge that no single task has perfect validity and have noted in the revised manuscript that future studies should aim to replicate our findings using alternative emotion understanding tasks (e.g., Cameron et al., 2022, *Child Development*).

16. L136 and following: Here, also, it would be helpful for the reader if you already shortly describe the two tasks that are used as the dependent variables of the regression models.

Response: We have now added a brief description of the two tasks as the dependent variables of the regression models according to your suggestion.

17. L191-192: "Indicating that their responses at the odd frequencies were stronger than at the base frequencies." ==> This statement is incorrect. Responses at the odd frequencies are larger than the average response of the 16 surrounding bins, but not necessarily larger than the base rate 6Hz responses.

Response: Thank you for pointing this out. We have revised it in the manuscript accordingly.

18. L195-205: Please, report full statistics of the models, including the specific post-hoc (corrected?) comparisons.

Response: As mentioned in our response to the Reviewer's comment #3, we have added the key statistical results to the main text for clarity. The full details of the models, including all effects and post-hoc analyses, are now comprehensively reported in the SI. This approach ensures transparency while maintaining focus in the main text.

19. L209-220 (contrasting upright vs inverted faces): Please, provide full statistics for this analysis. E.g. was there a group x condition (upright vs inverted) interaction?

Response: We have now provided this information accordingly in the main text and SI. Please see Page 3, Lines 62-80 in the SI.

20. L261 "maximum deviation parameter": Has this parameter been validated already? For instance, to what extent is it also determined by (more immature) motor coordination, which probably also correlates with age?

Response: Yes, the maximum deviation (MD) parameter has been validated in previous behavioral and neuroimaging research as an index of how much the unselected category response was co-activated during perception (Stolier & Freeman, 2017; Freeman et al., 2011; Brooks & Freeman, 2018). While it is possible that motor immaturity in younger

participants could influence MD, it is important to clarify that MD primarily reflects the trajectory of mouse movement as it relates to competing decision alternatives, rather than raw motor accuracy or speed. Furthermore, in our study, we ensured that all participants demonstrated the ability to complete the task without significant difficulty, suggesting that motor coordination was sufficient for the task across all age groups. Additionally, the task's within-subject design minimized the influence of motor coordination by requiring relatively simple, straightforward movements.

21. L274 "the reciprocal of the SNR": How was this crucial measure exactly calculated? Across what channels / ROIs / harmonics …?

Response: We have now clarified how the reciprocal of the SNR was calculated on Page 12, Lines 329-331.

22. Facial Action Coding System: I am a bit surprised that the authors still used the manual coding, whereas nowadays there exist well-validated computer vision software approaches that can perform the job (e.g. FaceReader, OpenFace, …).

Response: We may explore using automated software in future studies. The advantage of FACS lies in its well-established validity and reliability in identifying subtle facial muscle movements, though we acknowledge that manual coding with FACS is time-consuming, and our coder underwent rigorous professional training and certification to ensure accuracy. Additionally, many automated systems are trained on datasets with limited representation of East Asian faces, which may result in missed nuances or cultural variations in facial configurations.

23. L309-315: Please, provide full statistical information for the models on the adults, including statistical information on the perceptual discrimination models. Can the authors speculate why neural discrimination scores no longer predict behavioural expression discrimination in adults (e.g. may neural scores be at ceiling and lack variability to predict behaviour?)?

Response: Statistical information for the adult models is now provided. Yes, we also suspect adults may exhibit ceiling effects in their perceptual discrimination of facial configurations for emotions, with individual differences in the task being more likely driven by their conceptual knowledge. Please refer to our response to your previous comment # 9 for detailed explanation.

24. L493-499: The authors apply a new pre-processing method where they remove 10s epochs of data that do not show good enough quality. While the approach may be valuable, it may result in less data (and thus less robust responses) in the younger age groups. Can the authors present an overview of the number of data segments that were ultimately included for each of the age groups? See in this respect also my previous comment that the main analysis should be the one that is calculated on the 20s segments for all participants.

Response: As per the Reviewer's earlier suggestion, we have updated the analysis in the manuscript to use the first 20 seconds of data for all participants across age groups. We have now provided the number of epochs included in the final analysis for each age group. No difference was found among age groups.

25. L544: "… Pearson correlations between children's and adult's ratings …" ==> Where are these correlations reported?

Response: This reference was mistakenly left in the text, but we have now removed it. We had originally conducted

an analysis comparing Pearson correlations between children's and adults' ratings. However, as this analysis was not directly relevant to the main narrative of the manuscript, we decided to exclude it from the manuscript.

26. L565-566: Verify the order of the examples in this sentence. It does not seem to fit.

Response: We have revised the sentence to ensure that the examples accurately correspond to the scenario pictures and emotions.

27. Fig 4B, 4D: Better to change low level, middle level and high level of age to younger, middle and older age group, respectively.

Response: We have updated Fig. 4B and 4D to replace "low level," "middle level," and "high level of age" with "younger," "middle," and "older age group," respectively.

28. SI L124, paragraph 3.3: What do these KDEF stimuli come to do here?

Response: The KDEF stimuli were initially used to explore potential cultural differences in emotion understanding. However, as this analysis was beyond the scope of the main objectives of our manuscript, it was not included in the paper. We overlooked its presence in the supplementary materials, so we have now removed this section.

29. SI Fig. 5: Provide more information on the exact analysis approach. E.g., how were these correlations integrated across multiple emotion pairs?

Response: We have added detailed information about the exact analysis approach, including how correlations were integrated across multiple emotion pairs, directly into the figure caption of SI Fig. 5.

30. SI Fig. 6: Provide the units of the Y-axis for Fig A and B.

Response: We have now standardized the units for the Y-axis in both Fig A and Fig B of SI Fig. 6 (now referred to as SI Fig. 7).

31. SI Fig.12: Please, also display the results of the non-significant prediction of behavioural emotion judgements by neural perceptual discrimination.

Response: We have now included all the results in the SI Tables.

RESPONSE TO REVIEWERS' COMMENTS_R2

Reviewer #1 (Remarks to the Author):

The authors have produced a substantial revision that addresses most of my original concerns. I have just a few small points remaining.

Response: We appreciate the reviewer's positive feedback on our revision. Below, we address the remaining comments point by point..

1. The abstract makes it seem like there are several studies each with 120-140 Ps… could you rephrase to clarify that these tasks were all implemented on one sample?

Response: Thank you for this helpful comment. We have revised the abstract to clarify that all tasks were implemented within a single sample of children, with minor variation in final sample sizes due to task-specific exclusions.

Changes to manuscript:
Page 2, lines 32-33, 41-42: "This set of studies, conducted within the same group of children, examines …" and "Minor variations in final sample sizes were due to task-specific exclusions."

*2. I feel like my response to my prior point 5 is still not resolved… Do EEG data suggest that young kids differentiating negative emotion expressions from each other or not? The main text claims "the ability to discriminate facial configurations has already emerged by 5 to 6 years of age," but I don't think either the main text analyses nor Supplemental Tables 4A-4G present the specific result I'm curious about (do young kids distinguish *negative* expressions). This feels like an important test to run and an important result to be clear about, as it weighs on lots of debate in the field about child representations of emotion being primarily about valence vs. specific emotions.*

Response: Thank you for this important follow-up. To address your concern, we have added condition-level results to the manuscript, reporting EEG responses for each emotion pair across age groups. These results show that children in the 5–6 years group exhibited reliable neural discrimination (i.e., oddball responses) not only for positive–negative contrasts but also between pairs of negative expressions (e.g., fear-anger, fear-sadness, anger-sadness). Full results are provided in the SI (SI Table 4).

Changes to manuscript:
Page 8, lines 209-214: "Furthermore, when examining the oddball responses for each condition separately, the 5–6 years group also exhibited above-threshold neural discrimination (signal-to-noise ratio > 1) across different emotion pairs, including those involving negative facial configurations (i.e., fear–anger, fear–sadness, anger–sadness, SI Table 4). This demostrates that even 5-6 year-olds are capable of discriminating between alternating facial configurations not only across valence boundaries but also within the domain of negative emotions."

Reviewer #2 (Remarks to the Author):

My thanks to the authors for their attentive responses to my comments – many of which have been fully addressed. There are a few that I think could use some more clarification or elaboration.

Response: Thank you for your thoughtful feedback and kind words. We have carefully revised the manuscript to address your remaining comments and clarify our interpretations. Please see our responses below.

1. In my previous point 1c, I questioned whether valence-based distinctions could be conceptual as well as perceptual. In their response, the authors note that their definition of valence highlights that it "is not merely a perceptual process but represents an integration of conceptual understanding that develops through children's embodied and social experiences." This is lovely phrasing – can it be incorporated into the text? There are also references to valence as a 'basic emotional cue' or 'sensory information' in the discussion (pp 15-16) that feel contradictory to an idea of it as integrated conceptual understanding. Can the authors further harmonize their thinking?

Response: Thank you for this thoughtful comment. In line with the suggestion, we have added a clarification regarding the definition of valence in the introduction. Additionally, we have revised the discussion to remove language that framed valence as merely a basic sensory cue, and incorporated phrasing consistent with the view that valence reflects an integration of both perceptual and conceptual processes.

Changes to manuscript:

In the introduction, we added the following clarification:

Page 3, lines 69-72: "Even early-emerging emotion judgments based on valence are likely informed by conceptual knowledge shaped through children's embodied and social experiences, rather than arising solely from low-level perceptual processes."

In the discussion, we revised the following passage:

Page 16, lines 454-460: "This developmental change aligns with the theory of constructed emotion, which posits that early in life, children rely on broad affective cues such as valence, arousal, and physical signals to infer others' feelings. While perceptually salient, these cues are increasingly interpreted through conceptual knowledge as children accumulate more emotion-evoking experiences and undergo cognitive development (Barrett, 2017; Barrett et al., 2019; Sloutsky & Deng, 2019)."

Page 16, lines 466-469: "In sum, children's evolving conceptual frameworks of emotions enable them to interpret emotional cues in increasingly flexible and context-sensitive ways..."

2. As part of my previous point 2, I wondered if the authors could state why they included only one positive emotion. I did not see a response to this point and think it would be useful to specifically point out this aspect when mentioning future directions (p 18).

Response: As suggested, we have clarified in the revised manuscript that the inclusion of a single positive emotion was a methodological choice to reduce task complexity and ensure feasibility with young children. We have also noted this as a limitation to be addressed in future work.

Changes to manuscript:

Page 18, lines 508-511: "In the current study, we included only one positive emotion category in order to reduce task complexity and maintain feasibility for younger participants. Future studies should incorporate a broader range of both positive and negative emotion categories to conduct a more comprehensive analysis."

3. In response to my previous point 10, the authors helpfully clarified that, in Study 2, children's conceptual knowledge was not being assessed for how well it aligned with cultural norms. However, there is a comment in the concluding paragraph about "the importance of exposure to cultural norms related to emotion" that now seems out of place.

Response: Thank you for catching this. We have revised conclusion to avoid implying that alignment with cultural norms was directly assessed.

Changes to manuscript:
Page 18, lines 520-523: "This developmental change may be shaped in part by children's emotional experiences and their socialization into ways of expressing and interpreting emotion."

4. More generally, I have the impression that the proposed developmental pathway for emotion understanding, discussed on page 15 of the discussion, is less clear in the revised manuscript than in the original. This has partially to do, I imagine, with the revisions in response to my point 11 (regarding comparison of Study 2 and 3 results). There is, for me, an uneasy tension between children's concepts becoming more complex (i.e., overlapping) in their features and simultaneously more differentiated in their deployment. Can the authors further specify their thinking here (e.g., via examples)?

Response: We appreciate this opportunity to further clarify our thinking. In the revised discussion, we have clarified the proposed developmental pathway and expanded the discussion by adding examples to illustrate how conceptual complexity and behavioral differentiation can coexist.

Changes to manuscript:
Page 15-16, lines 441-450: "As discussed, findings from Study II reflect higher-order categorization, where children recognize shared features and contexts across negative emotions (e.g. anger and fear can co-occur in situations involving perceived threat), indicating a more abstract and interconnected conceptual structure. At the same time, the improved behavioral differentiation in the emotion judgment tasks indicates that with age, children become better at differentiating the emotional nuances of these configurations in real-world contexts – for instance, recognizing that fear typically involves withdrawal and uncertainty, whereas anger is often associated with approach and blame."

Reviewer #3 (Remarks to the Author):

I was asked to review the authors' responses and revisions to the comments from Reviewer 3. I confirm that the present comments can be shared with the authors.
Overall, I find that the authors responded exceedingly thoroughly and thoughtfully to the Reviewer's comments. They undertook several additional rounds of analyses and replaced/expanded many of the original results. They also took care to anchor their methodological decisions in the literature, especially regarding FPVS. I found their arguments compelling but was also happy to see them point to open questions that can be addressed by future research.

Response: We sincerely appreciate your careful evaluation of our responses and revisions, as well as your positive and encouraging feedback.

1. I have three small comments in response:
In response to Reviewer 3 comment 1, the authors marshal evidence to suggest that the findings from the conceptual knowledge task were not impacted by younger children's cognitive abilities – even though they acknowledge that older children are generally more advanced in this domain. I didn't see this thinking reflected in the manuscript (e.g., as a response to potential alternative explanation in the discussion). Is there room to add it?

Response: Thank you for this helpful comment. In response, we have added a statement in the discussion to acknowledge the potential influence of younger children's cognitive and linguistic abilities, while clarifying why these are unlikely to account for the observed effects.

Changes to manuscript:
Page 15, lines 433-436: "While differences in cognitive and linguistic abilities could potentially influence task performance in younger children, we consider it unlikely that such limitations were the primary drivers of the observed developmental pattern, given the low rate of skipped items and the consistency in conceptual structure across age groups."

2. Similarly, in response to Reviewer 3 comment 20, the authors clarify that "MD primarily reflects the trajectory of mouse movement as it relates to competing decision alternatives, rather than raw motor accuracy or speed". Can this be added to the methods section where this measure is introduced?

Response: We have added a sentence in the methods section to clarify the interpretation of the MD measure, as suggested.

Changes to manuscript:
Page 24, lines 706-708: "MD has been validated in previous research as an index of coactivation between competing response options during decision-making, rather than a measure of raw motor accuracy or speed (Freeman et al., 2011)."

3. Finally, in response to Reviewer 3 comment 10, the authors state that older children's integration across negative emotion concepts "does not imply reduced differentiation but rather a more interconnected conceptual framework". In line with my comments on this revision, I am wondering if this language – and some examples - can be added to the discussion to help clarify this point. It is an important one.

Response: Thank you for this important suggestion. We have incorporated the suggested language and examples into the revised discussion. This clarification also aligns with revisions made in response to Review's comment #4; please refer to our responses above for further detail.

FINAL COMMENTS

Reviewer #1 (Remarks to the Author):

The reviewers have addressed all my remaining concerns.

Reviewer #2 (Remarks to the Author):

I was asked to review the authors' responses and revisions to my own comments (Reviewer 2) as well as those from Reviewer 3. I confirm that the present comments can be shared with the authors.

My thanks again to the authors for their conscientious and articulate revisions to the manuscript. I have no further comments and look forward to seeing this paper in print.

Reviewer #2 (Remarks on code availability):

I have previous provided comments on the code.