

Every Book Tells a Value:

Emerging Values and Alignment in Children's Book Recommendations

Andreea-Luminița Mocanu



Every Book Tells a Value:

Emerging Values and Alignment in Children's Book Recommendations

by

Andreea-Luminița Mocanu

to obtain the degree of Master of Science

at the Delft University of Technology,

to be defended publicly on Thursday August 27, 2026 at 11:00 AM.

Student number: 5334241

Project duration: November 17, 2025 – August 27, 2026

Thesis committee: Dr. Maria Soledad Pera, TU Delft, Thesis advisor
Robin Ungruh, TU Delft, Daily co-supervisor
Dr. Cynthia Liem, TU Delft, External examiner

“Grown-ups never understand anything by themselves, and it is tiresome for children to be always and forever explaining things to them.”

— *Antoine de Saint-Exupéry, The Little Prince*

PREFACE

As I add these final words, the much-awaited moment has arrived—I finally get to write the last bit of my thesis and let the emotions flow. Somewhere between finding my place in this process, I heard someone say that undertaking a Master’s thesis is equal parts personal journey as it is a contribution to research. These months have certainly been the most formative time of my life so far, especially the final stretch. The best part is that it has allowed me to grow immensely, in ways I could not foresee, and prove to myself I can carry out an extensive project across multiple months.

This would not have been possible at all without the unwavering support of my daily supervisor, Robin, with whom I have been lucky enough to work for longer than the span of this thesis. Together, we took an idea I had in a seminar, brainstormed, and crafted what is presented here. I am grateful for how naturally we understood each other and thoroughly enjoyed our collaboration. Thank you for pushing me to reflect and for all the feedback given these past months! After many tries and different methods, I am proud to have landed on something I believe in.

I would also like to thank Sole, my thesis supervisor, who has given me invaluable feedback and advice during the important milestones in this process. Sole has been an inspiration to me since the first Information Retrieval lecture with her, where the idea of considering technology for under-represented groups of users, such as *children*, sparked my interest. I feel genuinely fortunate to have benefited from her guidance throughout this journey.

Last, but definitely not least, I want to thank everyone that has been cheering for me in the background. Those who believed in me even though they had no idea what I was actually working on. To my dear friends and family: you are the heartbeat of this work. I would absolutely not be achieving this today without you.

Thank you to my Mom and Dad who, despite not understanding a word of what I have written here, supported me unconditionally for the past six years away from home. Their love pushes me to be better every day, and I hope one day to match the dedication you have shown me. To my sister, for growing up alongside me and shaping many of the *values* I now hold. To my little niece, who has been a lovely distraction this summer, and who I have kept in mind when writing and reflecting about children. To my friends who have supported me everyday all the way from Romania, Iulia and Larisa, for listening to me rambling and doubting myself through this process. To Radu, Aditi and Sofia, for making me feel at home here and offering me the extra push, but also comfort, needed to complete this. And to my close friends who keep up with me everyday and everyone else that showed up to support me—*mulțumesc!*

Andreea-Luminița Mocanu
Delft, August 2026

ABSTRACT

Recommender Systems (RS) play a crucial role in shaping children’s digital experience, by selecting which media they interact with. Through this selection process, RS not only influence immediate choices but may also contribute to the formation of preferences and worldviews. As is the case for all technologies, RS embody implicit and explicit design choices that reflect particular values. In this paper, we investigate the extent to which RS reflect children’s values in the domain of books. Since books represent a popular medium of introducing moral concepts to children, we will focus on this domain, but our approach can be applied to other textual content as well. Given that children are in a formative developmental stage, their values are both distinct from those of adults and subject to change, making them especially susceptible to being influenced by RS. We first analyze children’s book reviews to capture how values are expressed in practice. Based on these value expressions, we construct value profiles and obtain user segments through clustering, representing groups of children with similar value priorities and expressions. We use this in our analysis to gain insight into emerging patterns. Multiple recommendation strategies are employed to examine how well recommended items align with users’ value profiles. By comparing the values reflected in recommended books with those expressed by users, we assess whether RS amplify, suppress, or remain neutral with respect to children’s values. Our findings provide insight into how children manifest value preferences online and into the value alignment of RS for young users.

CONTENTS

Preface	ii
Abstract	iii
1 Introduction	1
2 Background and Related Work	4
2.1 Recommender Systems as Value-Laden Systems	4
2.2 Children as a Distinct and Understudied Population	6
2.3 Modeling Heterogeneous Users: Personas and Segmentation	8
2.4 Detecting Values in Textual Content	8
3 Experimental Framework	11
3.1 Dataset	11
3.2 Measuring Value Prioritizations	13
3.2.1 A Text-based Approach	13
3.2.2 Creating Value Profiles.	14
3.2.3 User Segmentation	18
3.3 Probing Recommender Alignment	19
3.3.1 Recommendation Algorithms	19
3.3.2 Evaluation Protocol	20
3.3.3 Metrics	20
4 Results	23
4.1 Value Profiles	23
4.2 User Segmentation	24
4.3 Sub-clusters within User Segments	26
4.3.1 User Segment A — Moderate, Balanced Expressor	26
4.3.2 User Segment B — Disengaged, Conservation-Leaning Expressor	29
4.3.3 User Segment C — Engaged, Prosocial Expressor	30
4.3.4 User Segment D — Selective, Openness-to-Change Expressor	32
4.4 Evaluating Value Alignment of Children and Recommendations	32
4.4.1 Value coverage	32
4.4.2 Amplification and Suppression	33
4.4.3 Accuracy	38
5 Discussion	40
5.1 Implications for Children	44
5.2 Implications for RS Design	45
5.3 Limitations and Future Work	46
6 Conclusion	49
References	50
A Additional figures for results of value profiles	55

INTRODUCTION

People today face an unprecedented volume of information and media to consume online. To help navigate this space and access items that might be of interest to them, recommender systems (RS) curate items for their users [63]. These systems encompass algorithms that predict what items a user may like and make suggestions accordingly. Recommendations range from which movie to watch next, what product to purchase, and whose posts people see on their social media feeds. By deciding which items are surfaced and which remain unseen, RS increasingly shape the information and perspectives users encounter in everyday life.

The capacity to shape what users encounter does not arise by accident; it is the result of concrete design decisions regarding the algorithms used, the data from which they are built, and the objectives for which they are optimized. This trade-off between what is prioritized and what is left out is unavoidable, and through it, RS shape the experiences, opportunities, and choices available to users as they interact with a given platform [53]. Consequently, the design of RS always inherently involves some normative dimension. Whatever configuration of design, algorithm and data is chosen or is not chosen prioritizes certain outcomes over others. However, which outcomes are prioritized in this way and at whose expense is rarely explicit in RS research, but is a result of underlying data and algorithmic blackboxes.

Over time, these individual-level effects accumulate into something broader, shaping the ideas, beliefs, and worldviews reflected in what users repeatedly see. For example, a news recommender that consistently favors emotionally engaging content does not just change which articles someone reads on a given day. It can gradually narrow the range of perspectives, sources, or issues they come to see as normal or credible. These outcomes are not incidental, but rather follow from the principles a system uses to decide what counts as relevant or worth surfacing. In other words, RS reflect **values**.

Systems are not neutral in which values users are repeatedly exposed to, nor in whose values those systems ultimately reflect [39]. According to Schwartz [66], a value is defined as something people consider important in life, “desirable goals that motivate action” such as equality, safety, and achievement. When it comes to technology, these values can be implicitly embedded, passed down from designers and engineers, or explicitly, when designed to promote particular goals or norms [54, 58].

The values embedded in RS matter because these systems shape the information, perspectives, and worldviews users are exposed to. Recognizing this influence, RS research has started to move beyond simply optimizing for accuracy towards broader human values, by developing algorithms to support democratic values [38], well-being [44], or foster self-actualization [45]. Similar developments can be seen in the industry. Twitter introduced metrics for “healthy conversation” [85], while Facebook incorporated “meaningful interactions” into its recommendation objective to encourage more socially valuable engagement [75]. These endeavours align with the movement of Digital Humanism, which advocates the design of technology around human needs and values rather than purely technical objectives [91]. One popular approach within this movement is Value-Sensitive Design (VSD), which proposes systematically incorporating human values throughout the design process by considering the often competing priorities of different stakeholders [32]. Together, these developments shift the question from whether RS embody values to which values they should support, whose values they should reflect, and how these choices can be made transparently.

Thus, there is growing consensus that technology, and RS in particular, should explicitly acknowledge and incorporate human values rather than treating them as unintended by-products of system design

[34]. However, relatively little empirical work has examined whether RS even reflect the values of the people who use them. Recent work has proposed conceptual frameworks for integrating values into RS and identifying indicators to evaluate them [76]; these approaches primarily focus on how systems ought to be designed rather than investigating which values are currently represented through recommendation processes. Further, they typically focus on broad stakeholder groups, such as platforms, content producers, and consumers, rather than the diverse values expressed by individual users. Relying on generalized assumptions about users can inadvertently marginalize groups whose needs and experiences do not fit these defaults. As a result, although RS inevitably privilege certain values through the choices they make about what to surface and suppress, there is only a limited understanding of whose values are actually being reflected and whether these align with the values of the people using them. Before asking how RS should support users' values, it is therefore equally important to understand what values users express and whether current systems represent them.

We focus on children as a user group of RS, as they are highly susceptible to the influence of RS on their values. While RS are increasingly used by children across different domains, little is known about whether these systems reflect or reinforce the values that children themselves express. This question is especially relevant because children's values are still under formation, unlike those of adults, which are comparatively stable. Children are in a developmental stage during which their sense of self, including their values, is still taking shape. Since their identities are still developing, values offer insights not only into who children are now, but into who they may become as well [23, 40]. Longitudinal studies further show that children's value priorities (which values one considers more important than others) differ from those of adults [24, 18, 22]. RS designed for or used by children, therefore, cannot simply rely on assumptions derived from adult populations, making it important to understand children's values on their own and how they are represented in recommendations.

Moreover, given that children are in a formative stage of their life, their value priorities continue to develop throughout middle childhood (6 to 12 years) before becoming more stable in adulthood [21, 18]. Thus, repeated exposure to value-laden content can contribute to the broader environment in which children develop and foster beliefs and worldviews. Since much of children's online experience is mediated by RS, the values embedded in recommendations may be particularly important for this developmental stage, potentially amplifying some values while suppressing others. Addressing this gap first requires understanding what values children express and then capturing how these values are reflected in the recommendations they receive.

To address the lack of understanding of how RS reflect children's values, we center on the book domain. Books serve as an impactful use case because they are a primary medium for introducing and teaching moral principles [15]. From educational settings to leisure, books are introduced in children's lives from an early age. They can shape worldviews and present children with narratives and fictional characters that help them internalize values such as kindness, patience, or friendship, through the perspective of others [12, 64]. Protagonists can often help children navigate fundamental developmental conflicts, such as the balance between security and independence [10], helping them grasp abstract concepts. Given the formative role books play in children's development, the choices RS make about which books to surface carry more weight. This motivates the following research questions (RQs) for our study:

- *What values do children prioritize, as expressed through their book reviews?*
To address this RQ, we use the ten universal values defined by Schwartz [66], which is an established framework. We identify the extent to which values are expressed and how they are prioritized by children in textual content, namely book reviews. The manifestations of different values create value profiles that describe each child by the degree to which a value is expressed in their book reviews. To identify patterns of similar value expressions and capture their treatment in recommendations, we cluster children's value profiles to obtain user segments, allowing us to examine salient value preferences among children as well as salient differences between them.
- *To what extent do book recommendations represent children's individual values, and how do they compare in terms of reflecting, amplifying, or suppressing these values?*
To answer this, we identify values represented by the items (books) through the metadata provided (i.e., book descriptions), so that items' and users' values can be compared regarding what they reflect. We employ different RS strategies (from traditional to state-of-the-art) to obtain a list of recommended books for each user, allowing us to assess whether patterns of value representation

hold across different recommendation approaches. Then, we measure the alignment between the values we considered important for each user and the values we most salient for the recommended books. We compare and contrast how the RS accommodates different user segments, revealing whether some value profiles are better aligned with recommendations than others.

Our first contribution lies in the insights gained into how children manifest values in their book reviews, through systematic identification of their prioritized values. Secondly, we contribute with an empirical exploration of the degree to which RS amplify, suppress, or neutralize values for this domain. Society sees children as requiring protection and recognizing it as their right [31], indicating that their needs should be considered explicitly. Our effort goes towards fulfilling this call to look at children in isolation. Without explicit consideration of them in research, RS can narrow their opportunities for growth [48]. RS risk systematically disadvantaging young users by reinforcing adult stereotypes, presenting them with inappropriate content, or propagating agendas to which children are more susceptible. By grounding our analysis in children’s own expressed values rather than assumptions inherited from adult populations, our work provides an empirical basis for evaluating whether RS are responsive to children as a distinct user group. As a by-product, our results can inform the design of value-sensitive recommendations that better account for children’s needs. Our implementation is publicly available at <https://github.com/andreealmocanu/values-children-reviews>.

BACKGROUND AND RELATED WORK

2.1. Recommender Systems as Value-Laden Systems

Recommender systems are comprised of algorithms designed to suggest items that individual users are likely to find interesting or useful [59]. These systems facilitate and steer access to digital content across diverse platforms, including news sites, social media, and e-commerce sites. Traditionally, recommendation approaches are categorized into content-based filtering, which suggests items similar to a user’s previous choices, and collaborative filtering, which identifies patterns among users with similar preferences [63]. Modern strategies also combine both content-based information for initial suggestions to make up for limited user interaction data, and introduce diverse items to counteract over-specialization, resulting in a hybrid approach [1].

Any configuration of algorithm, data, and objective prioritizes some outcomes over others, so that whatever a system surfaces or withholds shapes the experiences, opportunities, and choices available to users on a given platform [53]. Specifically, repeated exposure to a particular selection of content can gradually narrow the perspectives users see as normal, what sources they consider credible, or which ideas are worth pursuing. Given these consequences, it is important to look at what these systems embody.

RS inherently embed and amplify certain viewpoints through the dataset they are trained on, their objective or ranking function used [37]. Systems therefore differ not only in how accurately they predict user preferences, but in whose preferences and priorities they end up surfacing. Thus, this curation shapes what users repeatedly encounter, rather than being neutral [39, 77]. This is increasingly reflected in how researchers evaluate these systems. Although historically technical optimization and predictive accuracy were prioritized, there is a growing shift toward “beyond-accuracy” objectives such as diversity, novelty, and serendipity [42, 59, 50], reflecting a broader recognition that accuracy alone does not capture what makes a recommendation good for a user or for society.

This shift is not merely conceptual. Concrete instances of this normative shift toward acknowledging the values underlying RS outcomes include the design of algorithms intended to support democratic discourse in news recommendation [38], to align recommendations with user well-being [44], or to foster self-actualization [45]. This reframes recommendation as a socio-technical tool whose objectives can be adapted relative to stakeholders and domains, rather than a purely technical optimization problem.

These efforts align with the broader movement of Digital Humanism, which calls for designing technology around human needs and values rather than purely technical objectives [91]. Applied specifically to recommenders, this literature argues that norms for RS should empower users through transparency and control over profiling and personalization, and that different application domains require different interpretations of the same underlying principle, such as choosing between privacy or autonomy in different contexts [61]. One influential approach within this movement is Value-Sensitive Design (VSD), which proposes accounting for the values of different, and often competing, stakeholders throughout developing the system, through iterative conceptual, empirical, and technical investigation [32, 76]. VSD provides a vocabulary for reasoning about competing stakeholders, such as users, engineers, platforms, and motivates why purely technical metrics are insufficient without prior conceptual clarity about which values are relevant and whose prioritization is being represented. Taken together, this body of work reflects a growing consensus that systems should explicitly incorporate human values rather than treat them as an unintended by-product of other design choices [34]. Yet acting on this consensus

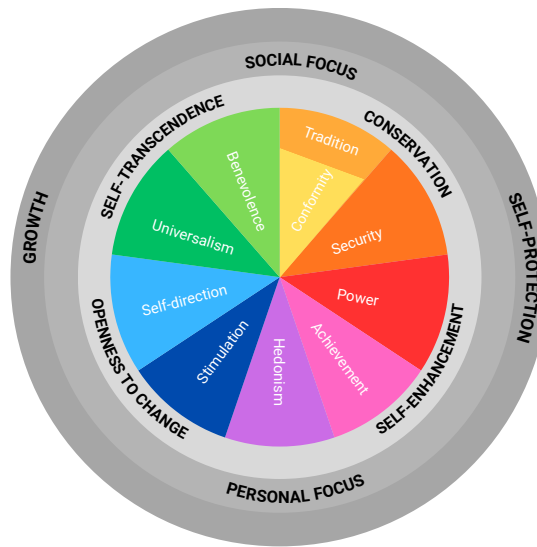


Figure 2.1: Schwartz’s value circumplex, as defined in [66].

requires a clear definition for what a “value” is and how individual values relate to one another.

Following Schwartz [66], from a psychological perspective, a human value is understood as a desirable, trans-situational goal that motivates action and serves as a guiding principle in a person’s life, such as security, achievement, or benevolence. The Schwartz Theory of Basic Values specifies a universal typology of ten basic values structured around motivational conflicts, such as openness to change opposing conservation (i.e., a preference for independent thought and new experiences versus a preference for stability, tradition, and order) and self-enhancement opposing self-transcendence (i.e., the pursuit of one’s own interest and success versus concern for the welfare of others and of nature)[66] (See Figure 2.1). The values are presented as forming a circumplex, which serves as the umbrella under which the majority of human values fall. For understanding what each value entails, we include their defining goal from Schwartz [66] in Table 2.1.

An alternative is the Rokeach Value Survey, which distinguishes between terminal values (desirable end-states) and instrumental values (preferred modes of conduct) [14]. We adopt Schwartz’s framework directly, on the grounds that it is cross-culturally validated across different ages too.

Applied to RS, such values can be embedded implicitly or explicitly [54, 58]. Implicit values are not intentionally designed, but emerge from seemingly technical decisions such as data collection choices, prioritized users, or which problems are considered worth solving. These values can also be designed for explicitly, when a system is built to promote a particular norm [54, 58].

Implicit values can also go unnoticed, since they are not intentional and thus not accounted for. This results in instances where technical design choices can embed structural inequalities into a system, undermining values such as fairness in the absence of any deliberate decision to do so. For example, voice-assistants sometimes struggle with the accents and speech patterns of people from underrepresented groups [49], or classifiers that detect skin cancer from images of moles have been trained on datasets containing only a handful of images of Black skin out of hundreds of thousands [30]. Such cases do not necessarily reflect explicit prejudice, but they do reveal an implicit judgment about whose data, and by extension whose experience, is treated as the “standard” case. Systems that are used daily, such as search engines, can shape what users come to believe is true through the order in which results are presented or through the stereotypes reinforced by auto-completion; a 2013 United Nations campaign, for instance, drew attention to real Google autosuggestions completing “women should” with phrases such as “stay at home” or “be in the kitchen” [55]. Examples such as these illustrate how not considering

Table 2.1: Schwartz’s ten basic human values and their defining goals

Value	Defining goal
Self-Direction	Independent thought and action; choosing, creating, exploring
Stimulation	Excitement, novelty and challenge in life
Hedonism	Pleasure or sensuous gratification for oneself
Achievement	Personal success through demonstrating competence according to social standards
Power	Social status and prestige, control or dominance over people and resources
Security	Safety, harmony and stability of society, of relationships, and of self
Conformity	Restraint of actions, inclinations, and impulses likely to upset or harm others and violate social expectations or norms
Tradition	Respect, commitment, and acceptance of the customs and ideas that one’s culture or religion provides
Benevolence	Preserving and enhancing the welfare of those with whom one is in frequent personal contact
Universalism	Understanding, appreciation, tolerance, and protection for the welfare of all people and for nature

values explicitly can lead to undesired behaviors of technology.

An important distinction within the RS literature is that the “values” invoked in RS research are not always the kind of broad, psychological values studied in value theory. Diversity, novelty, or fairness, for instance, function more as instrumental system-level goals, which aim toward higher-order human ends, than as broader values, as defined by Schwartz [76]. This matters because hand-designed, instrumental metrics often fail to fully capture the complexity of the underlying human concept they are meant to represent, and optimizing for one such metric in isolation can unintentionally degrade others [77]. It is thus important to consider multiple values at once and how tensions between them can be resolved, as VSD literature also proposes [32].

This literature discusses primarily how RS should be designed, rather than which values current systems actually reflect. Recent work has proposed conceptual frameworks for integrating values into RS and for identifying indicators by which such values could be evaluated [76]. Such frameworks specify what a value-sensitive recommender ought to consider, but does not measure whether existing recommendations represent the values of the people who use them. They also tend to reason about values at the level of broad stakeholder groups, such as platforms, content producers, and consumers, rather than the values expressed by individual users, which risks relying on generalized assumptions that can obscure the needs of users who diverge from the default. As a result, although it is broadly accepted that RS inevitably privilege certain values, often implicitly, through what they surface and suppress, little empirical work has examined whose values are actually reflected in recommendations, and whether this matches what users themselves express.

2.2. Children as a Distinct and Understudied Population

The question of whose values are reflected in recommendations is particularly pressing for children. Values contribute to a person’s identity, and, because identity is still forming in childhood, are also indicative of who a child may become [23, 40]. Developmental and longitudinal research shows that children’s value priorities, which values they consider comparatively more or less important, differ systematically from those of adults [18, 22, 24], and that these priorities continue to develop throughout middle childhood before stabilizing later in life [21]. This has two major consequences for the problem of identifying whose values RS reflect and how they shape children who use them. First, value-aware RS research conducted with or for adult populations cannot be assumed to transfer to children, since the value structure that such systems would need to align with is different and still changing. Second, because children’s values are still forming, sustained exposure to a particular selection of value-laden content plays a larger role in shaping their developing worldview than it would for an adult with

already-settled priorities.

Research on children’s values has largely built on Schwartz’s theory of basic human values and its higher-order structure (self-transcendence versus self-enhancement, and openness to change versus conservation). Evidence suggests that the value structures and hierarchies found in adults already appear in childhood, including cross-cultural regularities and gender differences in value priorities [26]. At the same time, work on value development highlights that while children may endorse recognizable value priorities early on, their conceptualization of values becomes more abstract, trans-situational, and motivation-based with age, with a developmental shift particularly evident around middle childhood, roughly six to twelve years old [70].

Children are also not just one group of “small adults”. They are a heterogeneous group in ways that go beyond their age. Their identities are shaped by diverse social, cultural, and political contexts, resulting in differences along dimensions such as race, ethnicity, socioeconomic status, immigrant generation, relationship to disability or illness, and many more [2]. This diversity of background extends to value priorities, as children growing up in conflict-affected zones are found to prioritize security and tradition more strongly [24]. Family structure, such as whether a child has siblings, has also been associated with differences in value priorities [36], as only-children prioritize power and benevolence more. Despite this, how children’s implicit values can be identified in practice, and how this diversity is accounted for, remains largely underexplored, particularly in the context of RS.

Furthermore, children are recognized as a population requiring explicit protection in digital environments [31]. RS that fail to consider them explicitly risk narrowing their opportunities by reinforcing adult-centric assumptions or exposing them to inappropriate content [48]. At the same time, addressing this cannot simply mean shifting responsibility onto children or their guardians. Work on community-level RS objectives shows that individual-level improvements, such as giving users more control or agency, may be insufficient to achieve good collective outcomes, since many users do not exercise this control in practice, and individually reasonable choices do not necessarily aggregate into socially good outcomes [74]. Applied to children specifically, this suggests that while children’s own agency matters, delegating responsibility for value-aligned outcomes to children, or to guardians alone, may place an undue burden on them without addressing the systemic role RS play in shaping what children encounter in the first place. This motivates examining value alignment at the level of the system, rather than assuming it can be resolved solely through the guardian’s control.

Broader work on ethical and age-appropriate AI reinforces this sentiment. Simple age categories are insufficient because intellectual ability, maturity, and experience vary widely among children, and child-centered AI should therefore incorporate children’s own perspectives, support their evolving autonomy, and consider the role of guardians [88]. The literature also identifies a recurring mismatch between high-level ethical principles and technical practice: child-facing systems are frequently evaluated primarily through technical performance metrics, while developmental well-being and children’s own understanding of the system remain under-measured. A closely related synthesis of age-appropriate AI principles explicitly includes recommendation among the AI behaviors it addresses, and identifies personalized content targeting as a particular risk during adolescence due to heightened susceptibility to persuasive or addictive design [89]. Taken together, this literature suggests that responsibly designing RS for children requires measuring not only whether a system performs well technically, but whether it respects and reflects the values of the children who use it.

Despite being frequent users of online platforms, children remain an underrepresented and overlooked population in RS research specifically [79]. Datasets rarely contain granular age information for users under 18, and children cannot be treated simply as “small adults”, as discussed above. This gap is widened by a structural constraint on data availability: because of the data-protection requirements owed to minors, datasets rarely contain verified demographic information that would allow children to be identified directly, limiting the scale and granularity of research that targets this population [35]. In the absence of such demographics, prior work has relied on inferring likely child users from behavioral or interaction signals as a proxy [71, 78, 90], an approach also adopted in recent RS research on children specifically [80]. As a result, most current understanding of how RS treat children is built on indirect signals rather than ground-truth data, underscoring the need for methods that can more directly capture children’s own values and experiences.

Books represent a domain in which value transmission to children is particularly explicit. They are a primary medium through which moral principles are introduced and taught to children [15], present narratives and characters through which values such as kindness, patience, or friendship can be internalized [12, 64], and can help children navigate developmental conflicts, such as the tension between security and independence, through protagonists [10].

Studies on children’s literature frequently examine its ideological content, noting that books often convey society’s expectations through explicit moral lessons [12, 43]. Content analyses of books for children reveal that values such as responsibility, love, helpfulness, and family unity recur frequently, while values such as peace, freedom, and independent thinking are comparatively underrepresented [15, 64]. A complementary perspective emphasizes books as central in early childhood settings and argues for culturally diverse books that expose children to different worldviews, linking book content to socialization and identity development [12]. Prior work has also identified fairness and representation concerns specific to this domain, such as the distribution of author gender in recommendation lists [27] which has potential effects on a child’s self-perception and view of social groups. Thus, recommendation over children’s books is not merely about satisfying book preference, but includes developmental and societal stakes, especially when recommendation influences exposure and diversity.

2.3. Modeling Heterogeneous Users: Personas and Segmentation

Understanding whose values are reflected in recommendations, and how, is important. If RS research wants to leverage this understanding to better serve children, we must be able to model value alignment at a level of granularity that reflects how those values vary across users. A population-level account of value expression or value alignment risks obscuring meaningful differences between users by averaging over a heterogeneous population. Work on news recommendation has shown that user populations are highly heterogeneous in their engagement behaviors, and that users in different segments exhibit substantial variation in recommender performance, rather than a uniform effect [84]. This motivates analyzing value alignment not only for the population as a whole, but also at the level of interpretable user groups that share similar characteristics.

The notion of grouping users into interpretable profiles is closely related to the concept of personas from interaction design and child-centric design, traditionally constructed through direct engagement with representative users. In child-centric design specifically, personas are treated as distinct from their adult counterparts: rather than being organized around task-oriented goals, child-personas are built around childhood developmental needs and abilities [4, 20]. Because directly interacting with child participants raises additional practical and ethical constraints, data-driven alternatives that derive persona-like groupings directly from existing interaction data, without requiring direct user contact, are particularly relevant here. One such approach constructs textual user stories from interaction data and metadata of items, embeds them, and clusters the resulting representations to obtain persona-like user groups, applied to the movie domain [56]. This establishes that persona-like groupings can be obtained from behavioral and textual traces alone, without requiring direct participant involvement, which is a necessary condition for applying a comparable approach to a population that cannot be easily surveyed, such as children.

To the extent this literature could be surveyed, no prior work was identified that combines recommender systems, children, and sub-population segments grounded in values. Existing work on RS and children instead suggests, indirectly, that children have diverse preferences, that recommenders may perform unevenly for underrepresented groups within this population, and that the values driving these differences remain largely unaddressed.

2.4. Detecting Values in Textual Content

Studying value alignment empirically requires a way of measuring the values expressed by users and represented by items. Traditional approaches to measuring values rely on self-report questionnaires such as the Schwartz Value Survey (SVS) [67] or Portrait Value Questionnaire (PVQ) [68], in which participants are asked directly to rate how important some value-terms are to them or how similar they are to some described behavior. Such methods are limited by depending on what participants believe to be true about themselves, which does not necessarily correspond to their authentic behavior or implicit preferences [13].

Comparing essays about personal values and behavior against self-reported questionnaire responses, prior work found that free-response text showed greater promise than questionnaires for capturing what people find important in everyday life [13]. An alternative is inferring values from naturally occurring text, such as reviews, rather than from elicited self-reports.

This is increasingly feasible at scale. The explosion of user-generated content, such as blogs, tweets, and product reviews, has provided an unprecedented opportunity to estimate human values from language [5, 47, 6]. Moral and personal values are deeply embedded in everyday language, and computational linguistics techniques allow for their automated extraction at scale [5, 72].

This general limitation of self-report is amplified for children specifically, for both cognitive and expressive reasons. On the cognitive side, children's structured self-reports are documented to be less reliable than adults' due to still-developing cognitive abilities [35], and direct, language-heavy instruments developed for adults can impose demands children struggle to meet. Studies assessing the Portrait Value Questionnaire (PVQ) in children report that while it can be used with older children, it is considerably more difficult for younger groups due to abstract language and the sustained attention it requires [22]. Child-centered alternatives, such as picture-based or multimedia instruments, have been developed to reduce this burden and focus on children's cognitive and developmental stage [25]. Even young children can make consistent value choices using such formats, though some deviations from Schwartz's circumplex ordering have been observed, particularly in the placement of tradition and conformity, suggesting that the mapping between adult values and children's own interpretations may not be perfect [19].

On the expressive side, children may not articulate their preferences in terms of abstract values at all: work with young children in user-experience interviews finds that cognitive constraints often lead children to justify choices via concrete, context-bound reasons rather than stable internal values, suggesting that methods for detecting values in children's natural language may need to infer values indirectly rather than rely on children naming them explicitly [93]. Complementing this psychological and HCI evidence, work on VSD for children argues for conceptualizing children's values in ways that reflect their lived contexts, such as home, family, or safety, and for participatory approaches in which children contribute to defining what matters, rather than only receiving adult-imposed value frameworks [29, 9].

Within data-driven approaches to value measurement more broadly, two main strategies can be distinguished. Dictionary-based approaches represent values as fixed lists of associated words and estimate the presence of a value from the frequency of matching terms in a text. For example, prior work has applied a hierarchical lexicon of fifty value concepts, constructed from social media posts, essays, and literary fiction, to user-generated content [72], building on earlier dictionary resources such as the Personal Values Dictionary [60]. This approach assumes that the frequency of value-laden terms correlates with the importance of the underlying value, but depends on exact matches against a fixed vocabulary that may not reflect the language of children since it was not constructed with this population in mind.

Embedding-based approaches instead represent values and text as vectors in a shared semantic space and estimate value expression through semantic similarity rather than lexical overlap, which allows value expression to be detected even when a text does not use the exact words contained in a fixed dictionary. Related work along these lines includes Distributed Dictionary Representation, which applies word embeddings to achieve semantic rather than literal matching [47]. This general principle underlies SIMON, a method originally developed to score text by its similarity to word embeddings representing sentiment categories [6], and has more recently been applied directly to value detection in user-generated content [72].

Another approach draws on large language models to identify value-laden statements in text and classify their relation to a value taxonomy [92]. A similar approach proposes automate prediction of moral language, for instance via moral foundations; this work shows that lexicons and embedding similarity can be used to predict normative dimensions of text more generally, but the moral foundations it targets are not equivalent to the broader, motivational sense of values [5]. In the context of e-commerce, psychographic analysis based on word use in reviews has been used to calculate user demand for diversity and mitigate market over-specialization [41].

Across all of these approaches, a shared methodological tension is that reducing an inherently contextual and cultural construct to a single similarity score or lexical match involves some degree of simplification of the value being measured. Recent conceptual work on the operationalization of public values in algorithmic systems describes this general pattern as “value narrowing”, where a broad understanding of a value is converted into something technically tractable while discussion of it continues to invoke its wider meaning [57]. This tension is relevant to any text-based value-measurement pipeline, independent of which specific technique is used. At the same time, organic, self-generated text has been shown to carry a value structure that is internally consistent with established value theory [13, 16], lending some support that these methods can provide a reliable and viable foundation for detecting human’s values.

EXPERIMENTAL FRAMEWORK

In this section, we present the experimental framework used to address our two research questions: ‘what values children prioritize in their book reviews’ (RQ1), and ‘to what extent recommended books reflect, amplify, or suppress these values’ (RQ2). We do this in order to shed light on whether children’s recommendations are aligned with what they themselves value. For that, we use textual reviews and book descriptions as proxies for larger values reflecting users and books’ full content. The experimental set-up consists of two main stages. First, we **identify values** important to children, as expressed by their book reviews. We use embeddings to encode both reviews and value definitions, enabling us to detect values in user-generated content. To capture differences in value preferences and identify patterns in how children prioritize values, we extract user segments and analyze value expressions within these segments. These segments are obtained by clustering user’s value profiles. Considering user segments allows us to get an overview of children’s value preferences and examine how values are prioritized across groups of children, revealing whether distinct profiles of value preferences emerge. By grounding user segments in expressed values rather than demographics or interests, we capture differences in which values characterize different groups of users, shedding light on how children’s value preferences may vary within the user population.

Second, we measure **alignment** between users’ expressed values and those represented in books recommended by different recommendation strategies, to explore whether recommendations reflect children’s value preferences. For each user, a set of book recommendations is generated using a common recommendation algorithm. The values expressed or implied in the book descriptions are then combined to get one representation of the “recommended values” for each user. This allows us to directly compare how much the recommendation aligns with the expressed user values. To look at systematic (mis)alignment patterns, we center our analysis on user segments.

We start by presenting the dataset and pre-processing steps (Section 3.1) taken in order to apply the methods for identifying values in text, described in Section 3.2.1. Next, in Section 3.2.2, we outline the clustering strategy used for creating user segments. Finally, we present the different recommender strategies considered in this experiment in Section 3.3.1 and discuss the evaluation of value alignment between children’s expressed values and values represented in book recommendations in Section 3.3.2.

3.1. Dataset

In this subsection, we present the general set-up of the dataset and all pre-processing steps we take.

To analyze value alignment of recommender systems in an offline setting while remaining real-world-attuned, we use a well-established dataset containing extensive user-book interaction data. This dataset reflects a diverse user population with a broad range of user behaviors and preferences. By favouring a dataset containing broad realistic interactions over smaller, less rich child-centered ones, we ensure that our experimentation is based on rich user and item information.

We use the publicly available dataset from **Goodreads**¹ [86, 87]. The dataset was collected in 2017 by scraping Goodreads and provides us with metadata about books, user-book interactions, and book reviews written by users. In total, the dataset contains 2,360,655 books, 876,145 users and 228,648,342 user-book interactions (ratings and reads) and approximately 15 million book reviews.

¹Dataset publicly available at: <https://cseweb.ucsd.edu/~jmcauley/datasets/goodreads.html>

Table 3.1: Distribution of books, interactions, reviews and users across main genres in the Goodreads dataset.

Genre	Nr. of Books	Nr. of Interactions	Nr. of Reviews	Nr. Users
Children	124,082	10,059,349	734,640	92,667
Comics & Graphic	89,411	7,347,630	542,338	59,347
Fantasy & Paranormal	258,585	55,397,550	3,424,641	256,088
History & Biography	302,935	31,479,229	2,066,193	238,450
Mystery, Thriller & Crime	219,235	24,799,896	1,849,236	203,655
Poetry	36,514	2,734,350	154,555	47,400
Romance	335,449	42,792,856	3,565,378	198,141
Young Adult	93,398	34,919,254	2,389,900	209,152

The Goodreads dataset provides a large-scale representation of how readers interact with books in an online community. Users are represented through their reading histories and written reviews, while items are characterized by metadata on the content of the book (in the form of a description), on the author, and assigned genre. We consider the books a user has interacted with as their **user profile**. The user profile reflects readers’ opinions of books and their consumption behavior of books. For example, this allows us to identify users’ preferred genres based on their interaction history, as various books are assigned to one of 8 main genres in the dataset. Books assigned the *children* genre alone account for 8.5% of all books and 4.8% of all interactions. The distribution of available data across genres is presented in Table 3.1.

User reviews. In the book domain, user perspectives can be expressed through user-generated content in the form of reviews. Reviews are especially meaningful as we not only have access to books read or liked, but also to a form of self-expression that allows us to observe potentially underlying values. After reading a book, children can share their opinions, describing what they enjoyed, what they found troubling, or which aspects of the story resonated with them. For instance, children may express how they feel about moral situations depicted in the book or whether they react strongly to themes such as friendship, fairness, or courage. As a result, reviews can provide a rich source for identifying values, especially as they are written from the users’ own perspectives and experiences. Considering this benefit, our analysis relies on the assumption that values expressed in user reviews reflect broader values of the users.

Item Descriptions. Regarding the items, books are represented through their descriptions, which serve as the main form of item content. Book descriptions typically summarize the plot, introduce key characters, and can highlight central themes or conflicts. In this way, descriptions can implicitly signal values emphasized in the book. While descriptions are more formally written and less subjective than reviews, they still provide meaningful textual content from which values can be inferred. Our study relies on the assumption that values in descriptions serve as representative for the broader values expressed in the full book content. Consequently, the availability of book descriptions is essential for our exploration, as our analysis requires textual content for both users (reviews) and items (descriptions) as proxies for the users’ perspectives and for the books’ full content.

Preprocessing. The dataset contains 1,143,241,710 user-book interactions. For our experiment, we apply the following pre-processing and filtering steps:

1. We retain only interactions for which users wrote a review, reducing the dataset to 81,095,745 interactions. Reviews enable our analysis because they form the textual basis for inferring user values. Interactions without reviews only contain implicit feedback signals, such as ratings or reading activity, which do not provide semantic information about users’ beliefs or preferences. This filtering substantially reduces the dataset in size and impacts the interaction histories of users. Rather than feeding the recommender the full history, we are only keeping books that users felt motivated enough to review. However, this step is necessary for our exploration as reviews represent deliberate and expressive forms of engagement and provide us with textual content to understand the user. A sufficiently large corpus remains with large user profiles based on interactions where users felt strongly about the books.
2. Next, we binarize the items, in line with common practices [81, 83]. This means we only keep

interactions where users assigned ratings higher or equal to 4, resulting in 54,180,660 interactions. We focus only on books they liked, primarily because modern recommender systems rely on positive/negative feedback signals rather than rating scales. Additionally, this filtering step reduces the dataset to a more computationally manageable size for the recommender experiment.

3. Finally, we remove users with fewer than 5 remaining interactions, leaving us with 31,196,529 user-book interactions. This step improves the reliability of inferred user values and priorities, as users with few interactions provide insufficient information to characterize their preferences.

These steps facilitate our exploration as they ensure that only books and users with sufficient interaction data are retained for reliable analysis. Afterwards, we set aside 10% of all interaction data per user for testing, i.e., the **held-out interactions** (test set). The remaining interactions are treated as the **profile interactions** (train set), out of which 10% will be used for hyperparameter tuning (validation set).

Inferring Child Users. As the dataset does not provide demographic information, we rely on interaction behavior as a proxy for the sample that represents the target population—children. Considering children’s past behavior and preferences, we identify users who can be considered child users. This is a common approach in the absence of real demographics [71, 78, 80, 90]. Specifically, we deem child users as users who have consumed substantially more books within the *children* genre than the average user. We operationalize this by selecting users whose proportion of *children* books is at least 2 standard deviations above the population mean, rather than a fixed cut-off point. This results in a threshold of 22% of *children* books within a user’s interaction history.

This approach is motivated by the assumption that users with strong engagement in children’s literature are more likely to belong to, or closely represent, the target population of child readers. Therefore, books tagged with the *children* genre generally reflect literature intended for children, and might be predominantly consumed by children. While this does not guarantee that every interaction belongs to a child user, it provides a practical method for identifying child-oriented reading behavior in the absence of datasets that provide demographic information. As we are looking at users who have high interest in children’s books, this captures child-centric behavior. Applying this threshold results in 9,540 child users, representing around 4% of all interactions in the training data. This is in line with expectations reported in prior research on RS datasets across various domains [82].

After these pre-processing steps, the resulting dataset contains textual representation for both users and items: user reviews reflecting user perspectives and expressions, and book descriptions reflecting item content. Together with the interaction data, this provides a foundation for value detection in children’s book reviews and for studying value alignment between users and books in a recommendation setting.

3.2. Measuring Value Prioritizations

Our first research goal is to detect which values children express in their book reviews. To do so, we infer values from textual information. Specifically, we identify which values are expressed in user-generated content (book reviews written by children) and in item-related content (book descriptions).

3.2.1. A Text-based Approach

Traditional methods for measuring values and their priorities use self-report questionnaires. However, these methods are limited by relying on what participants think to be true and which questions researchers think of, and may therefore not capture authentic behavior or implicit preferences [13]. In real-life settings, we rarely ask users about their values. Moreover, in a study carried out by Boyd et al. [13], participants were asked to write short essays about their personal values and their behavior, which were later compared with self-reported questionnaires. Their findings suggest that self-report questionnaires may not paint an accurate picture of individual’s mental life, whereas free-response data show greater promise for capturing what people actually find important in everyday life. We therefore use data-driven tools to infer values from text, as this allows us to capture what users express as important in the course of natural interaction (in our case, reviews users write about books they read) without requiring setting up a study with children, which raise additional ethical concerns and restrictions.

To achieve this, we must represent both values and textual content (reviews or item descriptions) in a

comparable form to measure how closely a piece of text aligns with a particular value. One common method is to use a dictionary of values, which represents values as a list of words. For example, Shen, Wilson, and Mihalcea [72] use a pre-constructed hierarchical values lexicon containing 50 value concepts, constructed from social media posts, essays, and literary fiction. This approach works on the assumption that the frequency of a specific value-laden term in natural language correlates with the importance of the underlying value expressed in text. However, this approach relies on exact wording matches against a fixed vocabulary of terms. Since this lexicon was not specifically designed or adapted for children’s language, it may not adequately capture the vocabulary and expressions commonly used in children’s book reviews.

Our Approach. Instead of relying on a strict word corpus, our approach captures semantic meaning. Word and sentence embeddings provide such a representation by mapping text into a high-dimensional vector space in which semantically similar texts are positioned closer together. We create embeddings of value definitions to represent values quantitatively. Similarly, we embed sentences of reviews and descriptions to represent them in a high-dimensional space. This enables us to quantitatively compare value definitions with sentences extracted from reviews and descriptions through semantic similarity. We adapt this comparison from SIMON [6], a method originally designed to score text by its similarity to embedded words representing sentiment categories. In our case, the value definitions take the place of these word categories, so that a sentence’s similarity to a value’s embedding reflects how strongly that value is expressed. We provide a high-level overview of this approach in Figure 3.1, where we illustrate how we get a numerical representation for each user. With this approach, values are operationalized by translating abstract concepts into measurable proxies. Prior work has successfully applied embedding-based methods to value detection in natural language [72], motivating our approach.

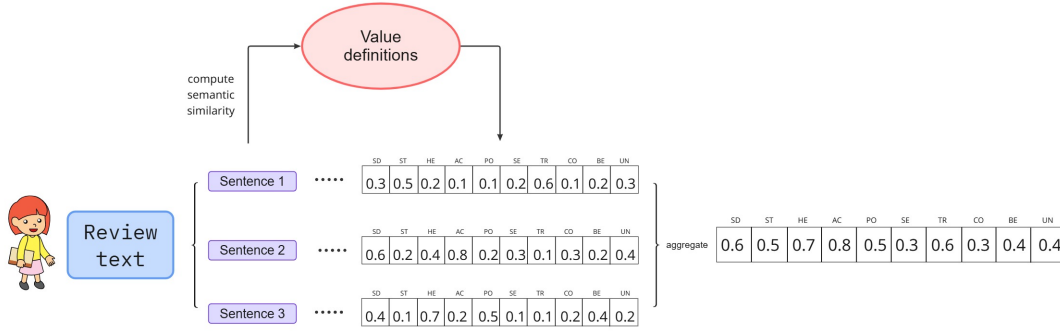


Figure 3.1: Value profile generation for each user.

3.2.2. Creating Value Profiles.

We create high-dimensional numerical representations of both values and reviews as vectors (embeddings) that capture their meaning, and then measure how similar they are to each other.

Value Embeddings. Our approach relies on values being represented uniquely in a high-dimensional semantic space. We explore two methods for constructing value representations. In the first method, each of Schwartz’s basic values is represented using its *textual definition* (Table 3.2). The definition is encoded directly into a sentence embedding, producing one vector representation per value. In the second approach, values are represented using the *value word dictionary* described in Section 3.2.1. Here, words associated with a value are embedded individually, after which the resulting vectors are averaged into a single embedding representing the value.

To determine the most suitable approach, we evaluate how well the resulting embedding spaces reproduce the theoretical value structure proposed by Schwartz [66], similar to [73] where the Rokeach value items are empirically validated against Schwartz’s. Specifically, values belonging to the same higher-order dimension should be located close to one another, while opposing dimensions should be separated. For example, openness to change should cluster apart from conservation, and self-enhancement should oppose self-transcendence, as described in Section 2.1. Following Schwartz’s original work [66], we use multidimensional scaling (MDS) to visualize the relationships between values.

For the approach based on *textual definitions*, we use a sentence embedding model, namely a pre-trained

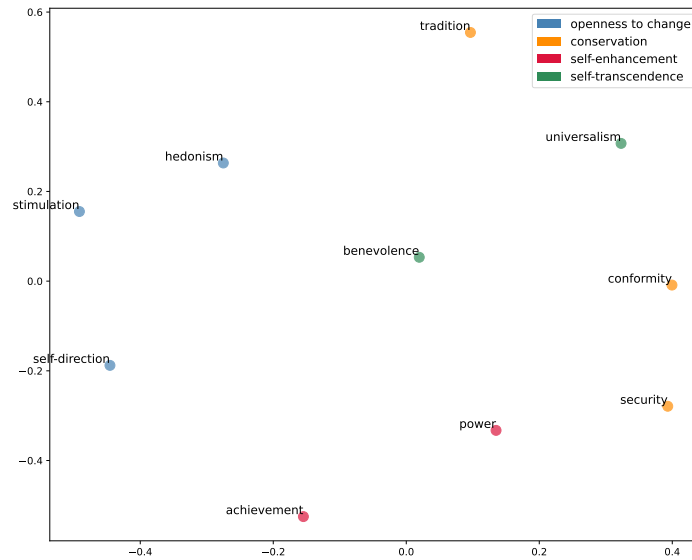
Table 3.2: Definitions of the Schwartz value types in the revised theory and abbreviations used (abbrv.).

Value	Abbvr.	Definition
Self-direction	SD	Independent thought and action, choosing, creating, and exploring
Stimulation	ST	Excitement, novelty, and challenge in life
Hedonism	HE	Pleasure and sensuous gratification for oneself
Achievement	AC	Personal success through demonstrating competence according to social standards
Power	PO	Social status and prestige, control or dominance over people and resources
Security	SE	Safety, harmony, and stability of society, relationships, and self
Tradition	TR	Respect of, commitment to, and acceptance of the customs and ideas that traditional culture or religion provides
Conformity	CO	Restraint of actions, inclinations, and impulses likely to upset or harm others or violate social expectations or norms
Benevolence	BE	Preservation and enhancement of the welfare of people with whom one is in frequent personal contact
Universalism	UN	Understanding, appreciation, tolerance, and protection for the welfare of all people and of nature

Note. Adapted from Sagiv and Schwartz (2022).

transformer model that maps sentences into a high-dimensional vector space. We explore both a smaller (MiniLM) and larger (mpnet) method for sentence embedding.

Figure 3.2 shows the values in a 2-dimensional space after applying MDS. The figure indicates that MiniLM largely preserves the expected value structure. Values belonging to the same higher-order dimensions cluster together, while opposing dimensions remain distinct. Although the conservation values are somewhat less cohesive, with tradition appearing further from conformity and security than expected, the overall structure remains consistent with Schwartz’s theory. In contrast, the larger MPNet model (Figure 3.3) produces a less coherent arrangement of values and weakens the separation between self-enhancement and self-transcendence. We therefore select MiniLM as the sentence embedding model for this approach.

**Figure 3.2:** Projection after embedding, using MiniLM and each value definition individually.

The other option consists of using a **value word dictionary** instead of value definitions [60]. In this case, we encode all the associated words per value using MiniLM and average the resulting vectors into a single representative embedding. This averaged vector serves as the point that summarizes the

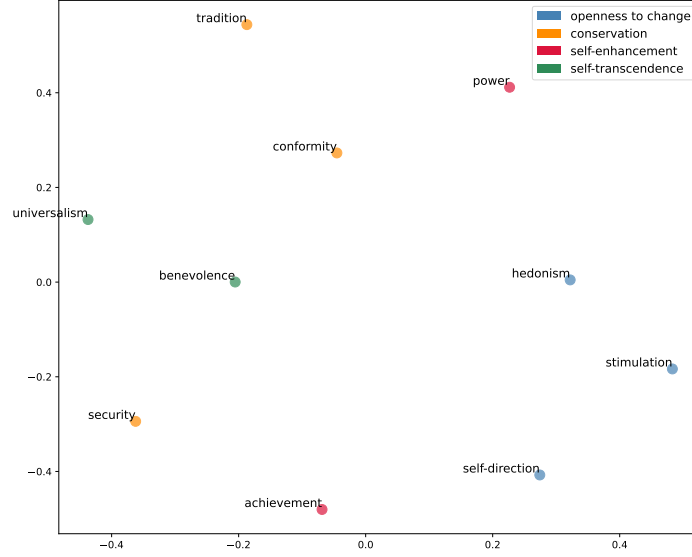


Figure 3.3: Projection after embedding, using mpnet and each value-word individually.

meaning of all words associated with that value.

We subsequently compare the definition-based approach with the dictionary-based approach. As shown in Figures 3.4 and 3.5, averaging embeddings of value-related words results in a substantially more compressed representation in which most values are projected close to one another. This limits discrimination between value categories and reduces alignment with Schwartz’s theoretical structure (see Section 2.1). By contrast, embeddings derived from value definitions reproduce the higher-order value dimensions more clearly and provide better separation between individual values.

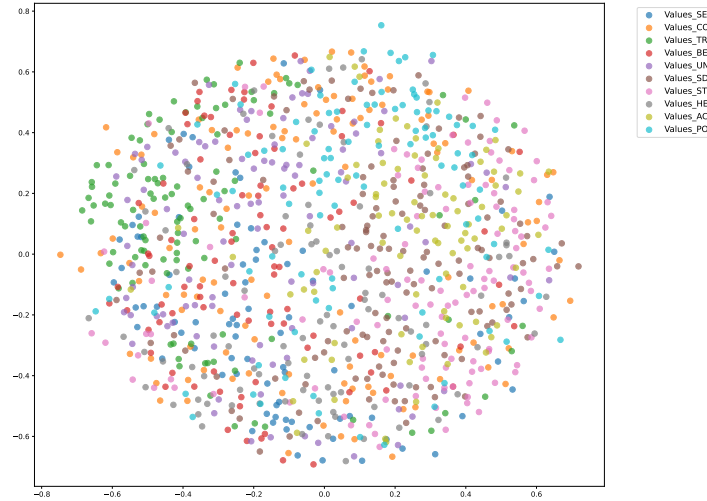


Figure 3.4: Projection after embedding each value-word individually to see whether natural clusters form within the same value.

Because our objective is to distinguish between different values when analyzing review content, we adopt the definition-based approach relying on sentence embeddings and use the MiniLM model for all subsequent analyses.

Creating Value Profiles. After identifying which embedding method to use to represent the values, we apply the same method to the reviews and obtain the **value profile** $\mathbf{v}_u \in [-1, 1]^K$ over Schwartz’s $K = 10$ basic values, where each component $v_{u,k}$ reflects the intensity with which value k is expressed in by user u from the set of all child users $\mathcal{U}$. Each book review written by a child included in **profile**

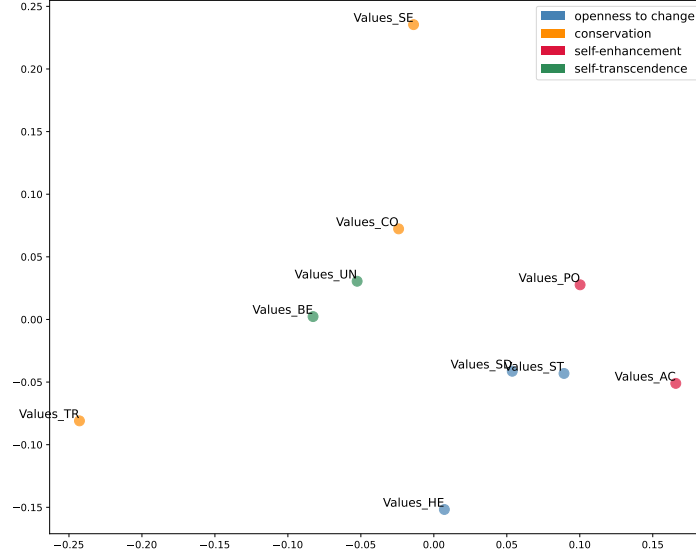


Figure 3.5: Projection after embedding each value-word individually, then taking the average for each value.

interactions is concatenated to form one piece of text, which is tokenized into sentences. We use the same model for embedding both values and reviews, ensuring that both are represented in the same vector space and are directly comparable.

To determine which values are expressed in a review, we compute the cosine similarity between each sentence embedding (vector for each sentence) and each of the ten value embeddings (per user). Cosine similarity measures the angle between two vectors, where a score of 1 indicates identical direction, meaning strong semantic alignment, while a score closer to 0 indicates little to no relation, and a score of -1 indicates opposite direction. To aggregate between different sentences, we take the maximum similarity per value, inspired by the approach from [72]. This maximum represents the degree to which each value is expressed by the user. If any of the sentences strongly expresses tradition as a value, for example, we consider tradition to be important to the user and take this score, rather than averaging over sentences, which could dilute this value expression. This is how we obtain a **value profile** $\mathbf{v}_u$ for each user, represented by a 10-dimensional vector containing the degree to which a value is expressed by the user, for each of Schwartz’s values.

For a better understanding of how RS may affect different groups of children, we group them into user segments to capture variation in their value preferences and engagement patterns, thus including both mainstream child users and groups of children who diverge from it. This approach follows recent work in news recommendation showing that user populations are highly heterogeneous in their engagement behaviors, and that clustering users into distinct segments can reveal how recommender performance varies substantially across groups rather than being uniform across an audience [84]. These segments provide an interpretable grouping of users that helps move beyond treating all children as a homogeneous audience by considering differences within the population. In the context of recommender systems, this is critical because recommendations are often optimized toward dominant behavioral patterns, which can obscure the needs and values of less represented users. By capturing distinct segments, we analyze whether recommendation outcomes differ across groups of children and examine how systems may reinforce or neglect certain value profiles.

Building on these foundations, we draw on the idea of personas from interaction design and child-centric design, while adapting the approach to support empirical evaluation of recommenders. Rather than constructing synthetic personas, we derive user segments directly from a dataset of existing users, without the need to interact with them directly. This allows us to use these segments as interpretable and measurable entities to assess whether RS provide appropriate recommendations for different user groups, based on their value profiles. We aim to find meaningful differences among users and go beyond mainstream patterns by looking at subgroups. Thus, we avoid overlooking users that are

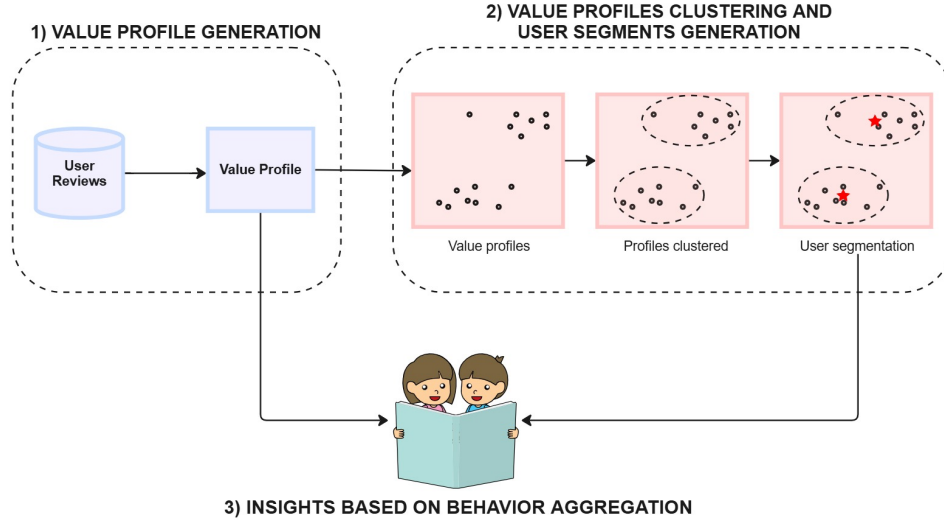


Figure 3.6: Methodology to create user segments from user review data, inspired from Misztal-Radecka and Indurkha [56]. Value profiles are clustered using K-means, which results in user segments. Based on their value profiles, we describe the segments to gain insights in step 3).

already underrepresented.

Approach. To identify user segments that reflect the values children prioritize in their book reviews and gain insights on patterns that may emerge, we adapt the approach from [56]. We present our approach in Figure 3.6, where we show how user segments are obtained at a high level. Originally applied to the movies domain, we adapt it for books, and move from personas to broader user segments. Misztal-Radecka and Indurkha [56] proposes constructing user stories by combining textual templates containing details about user-item interaction and their metadata, which are then embedded. Instead of user stories, we cluster users based on their value profiles, since they are already a numerical representation of each child.

3.2.3. User Segmentation

To identify groups of children with similar value prioritizations, we cluster users based on their *value profiles*. To prepare value profiles for clustering, we apply **z-score standardization** across each dimension independently, subtracting the dimension mean and dividing by the standard deviation. This step is taken to ensure that differences in baseline score levels across values are comparable and do not influence the clustering.

We apply a two-stage clustering procedure. First, we use K-means clustering to partition users into four broader groups², which we refer to as **user segments**. These segments capture the dominant patterns in children’s value profiles.

Subsequently, we apply HDBSCAN [51] separately within each user segment to identify more fine-grained groups of users. Prior to this step, we apply UMAP (Uniform Manifold Approximation and Projection) [52] to reduce the dimensionality of the 10-dimensional value profiles to two dimensions ($n_components = 2$; the other parameters are $n_neighbors = 20$, $min_dist = 0.0$ $random_state = 42$ for reproducibility). We apply UMAP prior to HDBSCAN, as its manifold-aware reduction has been shown to improve density-based clustering performance despite some known distortion of local density [56]. We use the resulting low-dimensional embeddings solely for clustering; all subsequent analyses are performed on the original 10-dimensional value profiles. Unlike K-means, HDBSCAN can detect clusters of arbitrary shape and is robust to noise, allowing it to uncover smaller and potentially less represented groups that may otherwise be overshadowed by the majority [56]. We refer to these groups as **sub-clusters** and use them as the basis for the subsequent analysis.

For each cluster, we compute the centroid by averaging the 10-dimensional value profiles of its members.

²The value $k = 4$ was selected based on the silhouette score and elbow method.

We interpret this centroid as the representative value profile of the corresponding user segment or sub-cluster, capturing the values that characterize the group as a whole.

3.3. Probing Recommender Alignment

Our framework focuses on the empirical exploration of *value alignment*. By value alignment, we refer to the extent to which recommended items reflect the values a child has expressed throughout their reviews, and whether this alignment holds consistently across the full spectrum of values rather than for only a subset of it. Concretely, for each user, we compare the value profile expressed in their reviews against the values expressed in the item descriptions of recommended books, and quantify this (mis)match for all value dimensions. This is illustrated in Figure 3.7, which shows our evaluation pipeline for RS and one user segment. We complement this with a traditional accuracy metric to examine whether value-aligned recommendations come at a measurable cost to conventional recommendation performance. Together, this framework allows us to move beyond a single aggregate accuracy score and instead ask *who* is well served by a given recommender, *which* values are over- or under-represented, and whether accuracy and alignment are in tension.

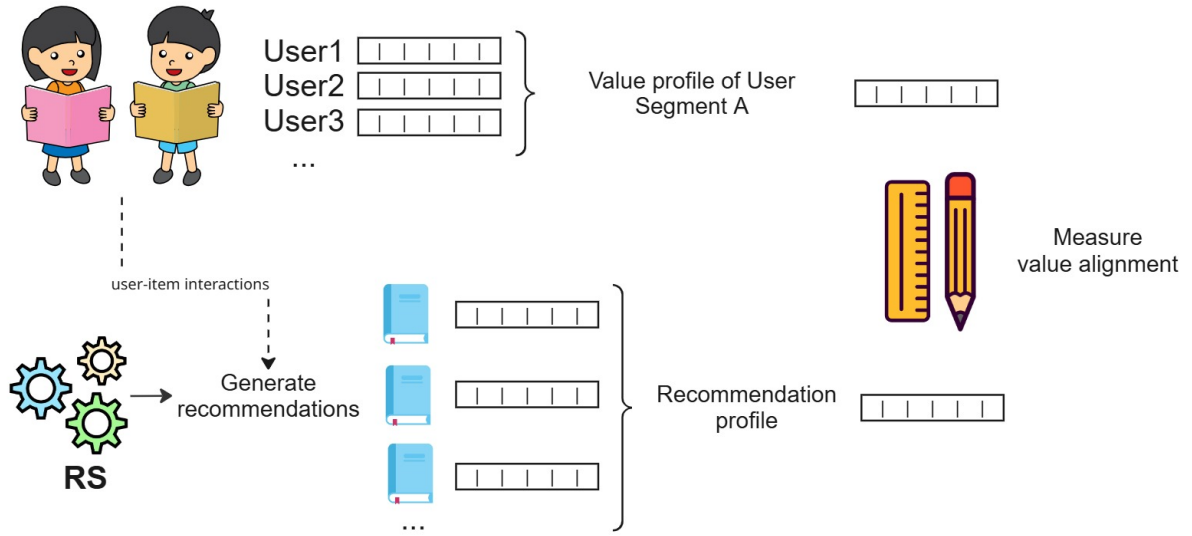


Figure 3.7: Evaluation pipeline, per user segment. Given the value profiles of the children from Segment A, we average to obtain the *Value profile of User Segment A*. Using different RS algorithms and the user-item interactions, we generate a list of recommended books. For each book, we also have the value profile, which we aggregate to obtain a *Recommendation profile*. Finally, we measure the alignment between these two profiles.

To ground this analysis, recall the data we described in Section 3.1. The dataset was split into a train set and a test set of user-item interactions, each annotated with the user reviews or item descriptions from which value expressions are extracted. The train set is used to tune the recommender algorithms and derive value profiles of children, while the test set is held out to evaluate the recommendations produced for each user. Building on the analysis from Section 3.2.2, we have already derived a *value profile* for each child from their reviews: a vector over the value dimensions of Schwartz’s basic values, representing intensity of each value as expressed in that user’s reviews.

3.3.1. Recommendation Algorithms

We consider three collaborative-filtering algorithms, chosen to represent well-established strategies in research [46]. *Most-Popular* serves as a non-personalized baseline, recommending the same ranked list of the most frequently interacted-with items to every user, after filtering out items the user has already interacted with. User-based k-Nearest-Neighbor (*USERKNN* [62]) recommends items by identifying the k users most similar to a target user (based on their interaction history) and recommending items those similar users engaged with, weighted by similarity. Item-based k-Nearest-Neighbor (*ITEMKNN* [65]) instead computes similarity between items based on co-occurrence in users’ interaction histories, and recommends items similar to those a target user has already interacted with. *RP3beta* ([17]) is a graph-

based method that models user-item interactions as a bipartite graph and computes recommendation scores via a three-step random walk, re-weighted by item popularity to counteract popularity bias. Together, this represents a meaningful range of personalization strategies.

Hyperparameter Tuning Before generating the recommendations, we run hyperparameter tuning using the Elliot framework [3] to ensure that each algorithm is evaluated at its best-performing configuration rather than under default or arbitrarily chosen settings. Tuning is performed by optimizing $nDCG@10$ on a validation set. We vary the number of neighbors uniformly between 5 and 1000 and the similarity to cosine and jaccard for ItemKNN and UserKNN. For RP3beta we vary the neighborhood value using a quantized uniform distribution between 5 and 1000 with a step size of 1, alpha and beta uniformly between 0 and 2, with a choice of cosine or jaccard similarity and switch between normalizing it. As a result of tuning, the best performing parameters are:

- **ItemKNN:** $k = 152$; similarity = *jaccard*.
- **UserKNN:** $k = 496$; similarity = *cosine*.
- **RP3beta:** neighborhood = 803; alpha = 0.451734; beta = 0.182912; normalize similarity = True.

3.3.2. Evaluation Protocol

This section describes how recommendations are generated and compared against a user’s expressed values, and introduces the metrics used to quantify value alignment and accuracy. We first describe how per-user recommendation profiles, which represent the values of the recommended books, are constructed, before turning to the alignment and accuracy metrics used to evaluate them.

Recommendation Strategy For each user u , the recommendation algorithms generate a ranked list of $N = 10$ recommended items based on the training set, $R_u = (i_1, \dots, i_N)$, using the Elliot framework. Recall from Section 3.2.2 that each user is associated with a *value profile* $\mathbf{v}_u$. We construct an analogous value profile for each book $\mathbf{b}_i \in [-1, 1]^K$ for each item $i \in \mathcal{I}$, following the same procedure but applied to book descriptions, such that each component $b_{i,k}$ reflects the intensity with which item i expresses value k . To compare a user’s expressed values against what they were recommended, we aggregate the value profiles of all items in R_u into a single *recommendation profile* $\hat{\mathbf{v}}_u \in [-1, 1]^K$ per user:

$$\hat{\mathbf{v}}_u = \frac{1}{N} \sum_{i \in R_u} \mathbf{b}_i$$

This recommendation profile $\hat{\mathbf{v}}_u$ is directly comparable in structure to the user’s own value profile $\mathbf{v}_u$, and forms the basis for all subsequent alignment metrics: rather than comparing a user’s profile to individual recommended items one at a time, we compare it to this single aggregated profile, which represents the overall value composition of everything a user was recommended.

3.3.3. Metrics

Here, we present the metrics used in our exploration. Since our alignment metrics rely on log-ratios between a user’s expressed and recommended value profile, and cosine-similarity-based scores in $[-1, 1]^K$ can be zero or negative, we first shift both profiles by the smallest value observed across all profiles so that every value becomes positive. We then normalize each profile so that its values sum up to 1, giving each value category a *share* of the user’s overall profile. This shift-and-normalize procedure is applied consistently to both value profiles and recommended profiles, using the same global minimum, such that the resulting shares remain directly comparable across users and profiles.

Coverage & Skew. We assess whether recommenders expose the full spectrum of values represented in the catalog, or whether recommendations systematically concentrate on a narrow subset of values regardless of what any individual user expressed. For each algorithm, we aggregate recommendation profiles across all users, weighted by how frequently each item was actually recommended, to obtain a population-level value distribution. We quantify how uniformly this distribution is spread across value

categories using *coverage*, the normalized Shannon entropy of $\bar{v}_k$:

$$\text{Coverage} = \frac{-\sum_{k=1}^K \hat{v}_k \log_2 \hat{v}_k}{\log_2 K}$$

Coverage is bounded in $[0, 1]$, with 1 indicating a perfectly uniform distribution across all K value categories, and values closer to 0 indicating that recommendation weight is concentrated on few categories.

As a complementary measure, we report *skew*, which captures the imbalance between the most- and least- represented value category in a recommendation profile $\bar{v}_k$:

$$\text{Skew} = \frac{\max_k \hat{v}_k}{\min_k \hat{v}_k}$$

Amplification. To identify *which* specific values drive any mismatch between a user’s expressed and recommended values, and whether that mismatch reflects over- or under-representation, we compute the $\log_2$ ratio between each user’s recommendation profile and their expressed value profile, per value category:

$$\text{Amplification}(k, u) = \log_2 \left(\frac{\hat{v}_{u,k}}{\mathbf{v}_{u,k}} \right)$$

A positive value indicates that value k is over-represented in a user’s recommendations relative to what they expressed (amplification), while a negative value indicates under-representation (suppression). We adopt a log-ratio formulation rather than a raw difference because differences of equal absolute size can represent different relative changes depending on the baseline value. For instance, a shift from a 0.05 to 0.06 share represents a 20% relative increase, while a shift from 0.40 to 0.41 represents only a 2.5% relative increase, despite both having an identical absolute difference of 0.01. The log-ratio formulation captures this distinction directly, and is symmetric with respect to direction: a doubling and a halving of a value’s share produce log-ratios of equal magnitude and opposite sign.

Distortion. While amplification identifies the direction and location of mismatch across individual value categories, it does not summarize the overall magnitude of (mis)alignment for a user. To capture this, we quantify individual-level value alignment using *distortion*, computed as the sum of squared per-category amplification scores:

$$\text{Distortion}(u) = \sum_{k=1}^K \text{Amplification}(k, u)^2 = \sum_{k=1}^K \left(\log_2 \left(\frac{\hat{v}_{u,k}}{\mathbf{v}_{u,k}} \right) \right)^2$$

This formulation is analogous to a mean-squared-error, but computed over log-ratios rather than raw differences, for the same reason we adopt a log-ratio formulation for amplification: it treats proportional over- and under-representation symmetrically, regardless of a value’s baseline share. Distortion is 0 when a user’s recommendation profile exactly matches their expressed value profile across all categories, and increases without bound as recommendations diverge from expressed values, whether through over- or under-representation of individual values.

Performance. To assess whether value-aligned recommendations come at the expense of traditional recommendation accuracy, we report $nDCG@10$, a standard ranking-based metric using binary relevance against each user’s held-out interactions. We examine the relationship between accuracy and our value-alignment metrics (amplification and distortion magnitude) using Spearman rank correlation, chosen for its robustness to sparse data.

Analysis For each metric, we aggregate the per-user alignment metrics within each user segment. This segment-level aggregation lets us move from a purely individual-level analysis to a structured account of *which groups of users* are served by a given recommender: for instance, whether children with a strong,

concentrated value profile receive more consistently aligned recommendations than those with a broad, diffuse one, or whether a specific user segment is disproportionately exposed to amplification of a particular value. Framing alignment this way allows us to describe patterns of mismatch in terms of *who* is affected and *why*, rather than reporting a single population-level number that could hide substantially different experiences across user groups. We provide a structured way to examine how different RS enact values across different user segments and to identify patterns of alignment or mismatch, allowing for more interpretability.

RESULTS

In this chapter, we present the results obtained from carrying out the experiments described in Chapter 3. In Section 4.1, we present the general characteristics of the value profiles from our dataset. Furthermore, we discuss the obtained user segments in Section 4.2 and the sub-clusters within each in Section 4.3. Finally, in Section 4.4 we discuss the alignment between children’s value profiles and recommended books.

4.1. Value Profiles

The value profiles obtained for the 9,540 identified children are summarized in Table 4.1. Scores represent the degree to which each value is expressed by a child, corresponding to the similarity between a user’s written review and value definitions, as outlined in Section 3.2.2.

Table 4.1: Descriptive statistics of value dimensions

	self-direction	stimulation	hedonism	achievement	power	security	tradition	conformity	benevolence	universalism
count	9540	9540	9540	9540	9540	9540	9540	9540	9540	9540
mean	0.335060	0.342925	0.325177	0.253867	0.283136	0.275802	0.263311	0.277719	0.272504	0.296914
std	0.083159	0.091320	0.082691	0.098824	0.094719	0.083856	0.111074	0.105332	0.087097	0.089662
min	0.005533	-0.000386	-0.025866	-0.096810	-0.092995	-0.049609	-0.090085	-0.094500	-0.082576	-0.045822
25%	0.279169	0.281878	0.269185	0.182713	0.216543	0.215893	0.184861	0.206509	0.211433	0.235614
50%	0.334319	0.342020	0.322086	0.248245	0.277979	0.267500	0.250017	0.276222	0.273806	0.291899
75%	0.389893	0.402223	0.378995	0.321379	0.346368	0.330639	0.329766	0.348128	0.335351	0.358081
max	0.623649	0.668306	0.790007	0.637470	0.665729	0.630720	0.714858	0.669345	0.539421	0.657460

The distribution of scores across value dimensions shows that users tend to express values moderately (for more details, see Figure A.1 in the Appendix), irrespective of the dimension, with mean values between 0.2539 and 0.3429, as reported in Table 4.1. This narrow range of means suggests that, at the population level, no single value dominates. However, the distributions themselves are not uniform and reveal meaningful variations. For example, *stimulation* has the highest mean, with a score of 0.3429, yet it is spread across users: 25% of them exceed 0.4022 and another 25% score below 0.2819, indicating considerable individual differences.

The heatmap in Figure 4.1 illustrates that children differ in their value expression, as indicated by high contrasts between colors across columns. We see that values like *self-direction* and *stimulation* have a darker tone across rows in the heatmap than, for example, *achievement*, suggesting these are more pronounced across users. For others, such as *tradition*, we see more contrasting colors. This is supported by our numerical results, where *tradition* has the widest min-max range and highest standard deviation (0.11), achieving second highest maximum across value dimensions and one of the lowest minimums (Table 4.1). This appears to be the most polarizing value for these users.

Based on the preceding analysis, we find that children’s value expression, as captured through book reviews, is generally moderate and broadly distributed across Schwartz’s ten value dimensions. No single value dominates at the population level, yet substantial individual differences do exist, most visibly in values such as *stimulation* and *tradition*. These differences motivate a deeper analysis, which we explore in the following sections. <https://www.overleaf.com/project>

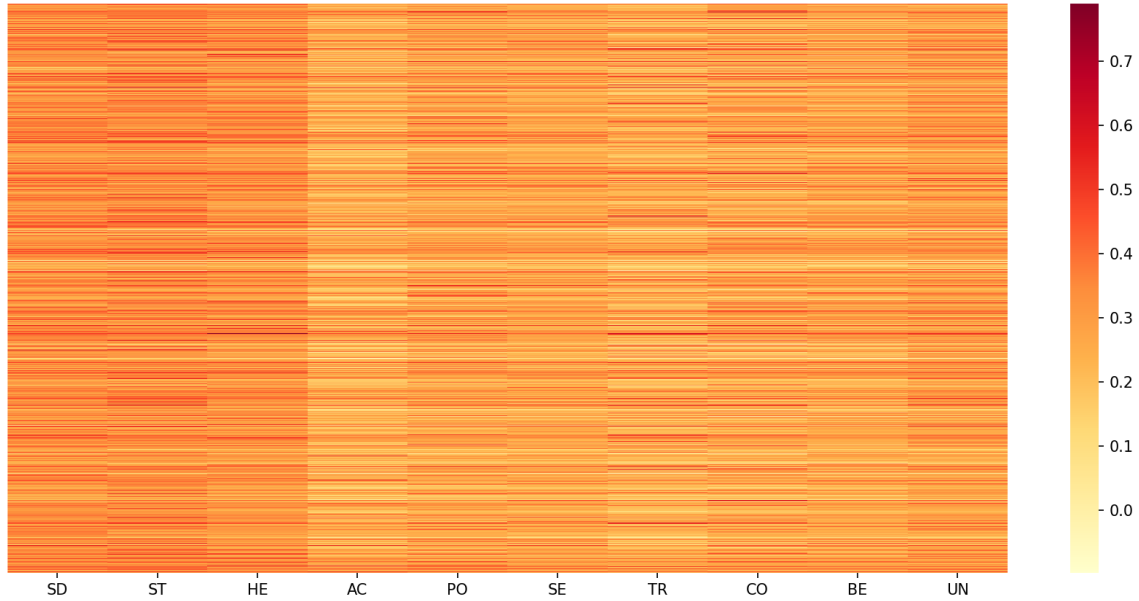


Figure 4.1: Heatmap of value scores for all 9,540 users.

4.2. User Segmentation

Based on the set-up described in Section 3.2.3, we derive user segments through a clustering approach. First, we apply k-means to group the 9,540 users based on their standardized¹ value profiles across 10 dimensions, as described in Section 3.2.2. The average standardized value preferences are visualized in Figure 4.2 and reported in Tables 4.2 and 4.3.

Table 4.2: Mean value scores for all user segments.

User Segment	self-direction	stimulation	hedonism	achievement	power	security	tradition	conformity	benevolence	universalism
A	0.364120	0.372387	0.354817	0.288410	0.317942	0.303697	0.297347	0.317184	0.309498	0.330533
B	0.213370	0.212198	0.206236	0.119557	0.148182	0.163970	0.120119	0.121404	0.139128	0.169050
C	0.438423	0.453391	0.429764	0.383482	0.410272	0.389818	0.411422	0.416140	0.385360	0.412472
D	0.300270	0.307531	0.288192	0.205280	0.235780	0.233118	0.209212	0.228130	0.229633	0.253891

Table 4.3: Mean value scores for all user segments, standardized per dimension.

User Segment	self-direction	stimulation	hedonism	achievement	power	security	tradition	conformity	benevolence	universalism
A	0.349466	0.322637	0.358460	0.349558	0.367489	0.332679	0.306441	0.374694	0.424762	0.374969
B	-1.463421	-1.431605	-1.438456	-1.359151	-1.424852	-1.333680	-1.289223	-1.484101	-1.531431	-1.426152
C	1.243011	1.209719	1.264846	1.311639	1.342319	1.359733	1.333511	1.314216	1.295815	1.288891
D	-0.418385	-0.387604	-0.447296	-0.491677	-0.499991	-0.509041	-0.487077	-0.470813	-0.492246	-0.479860

The four user segments mainly differ along two axes: overall **expressiveness level**, i.e., the degree to which children engage with the book overall, expressing their values, and **value preference**, i.e., which values they prioritize. Expressiveness is the primary distinguishing axis between the four segments, to which value preference is secondary and most informative when comparing segments with a similar expressiveness level (below or above average; a summary of these differences is reported in Table 4.4). Using these axes, we describe each user segment in detail below.

- **User Segment A — Moderate, Balanced Expressor** ($n = 3,240$; 33.96% of users): This is the second largest group. Value scores are consistently slightly above the population mean across all dimensions, with no single value standing out dramatically. *Benevolence* is expressed most ($z = 0.425$) and *tradition* least ($z = 0.306$), suggesting a mild suppression of conservation values,

¹After standardization, all value dimensions have a mean of 0 and a standard deviation of 1, with scores ranging from **-4.2454** to **5.6216** across dimensions.

though the difference is modest. The radar plot is broad and relatively flat, reflecting a group that engages with values evenly rather than selectively.

- **User Segment B — Disengaged, Conservation-Leaning Expressor** ($n = 1,248$; 13.08% of users): This is the smallest group and is defined by consistently below-average value expression across all dimensions, the lowest of all four segments. Standardized scores are strongly negative for all values (ranging from $z = -1.531$ for *benevolence* to $z = -1.289$ for *tradition*), indicating that this group's reviews contain the least value-laden language overall. Within this low level of expression, *tradition* and *security* are the least suppressed, giving the radar plot a distinctive relative lean toward conservation values, such as tradition, followed by security and achievement, although the difference is modest (roughly 0.24 standard deviations across all dimensions). However, looking at raw scores (Table 4.2) we see the segment B tends to have a mild preference for openness-to-change values: *self-direction* (0.213), *stimulation* (0.212), and *hedonism* (0.206) are the highest, thus the conservation lean should be understood as a relative tendency rather than a strong preference.
- **User Segment C — Engaged, Prosocial Expressor** ($n = 1,690$; 17.71% of users): This user segment represents the highest-scoring group across all dimensions, with broad, full coverage of the radar given by scores ranging from $z = 1.209$ for *stimulation* to $z = 1.360$ for *security*. Segment C expresses values the most intensely and broadly, but beyond that, *tradition* ($z = 1.334$), *security* ($z = 1.360$), and *power* ($z = 1.342$) are mildly more prominent, suggesting prosocial individuals with concern for self-protection. These users appear to engage with books through a lens of social order, care for others, and concern for stability.
- **User Segment D — Selective, Openness-to-Change Expressor** ($n = 3,360$; 35.22% of users): This is the largest group, with below-average value expression overall, but a relative peak in *stimulation* ($z = -0.388$), followed by *self-direction* and *hedonism*, which are least suppressed. Both in standardized and raw scores, openness-to-change values are most prominent for this user segment, while conservation and self-enhancement values are most suppressed, such as *security* ($z = -0.509$) and *power* ($z = -0.500$). Compared to the population, these individuals lean toward openness-to-change and away from self-enhancement values, valuing growth. These users engage with books through a lens of novelty and personal growth rather than social order or conformity.

To verify that these segments represent statistically distinct groups, we perform pairwise **t-tests** with **Bonferroni** correction across all cluster pairs and value dimensions. All pairwise mean differences are statistically significant ($p < 0.05$). To quantify the magnitude of these differences, we compute **Cohen's d** value for each pair (for more details, see Figure A.2 in Appendix). The largest separation is between **User Segment B** and **User Segment C**, with an effect size of approximately 4 standard deviations. This reflects that these two groups represent near-opposite ends of the value expressiveness spectrum. All other segment pairs differ by at least 1 standard deviation, confirming that the personas are clearly distinguishable from one another and represent broader trends within the dataset.

Segment B and D stand out with distinct shapes that show relative peaks, expressing one higher-order dimension more than others. Both are below-average expressors, but pointing in opposite directions (Figure 4.2): Segment B leans towards conservation-related values such as tradition and security, while D shows a clear spike toward stimulation, prioritizing openness-to-change. These two segments stand out compared to A and C, where the radar plots resemble a more circular shape.

Table 4.4: High-level description of each user segment identified by k-means clustering.

User Segment	Expressiveness Level	Value Preference
A	above average	balanced across dimensions, mild suppression of Tradition
B	way below average	relative peak towards Tradition, Security, Achievement
C	way above average	balanced across dimensions, mild inclination towards Self-protection
D	below average	relative peak towards Stimulation, Self-direction, Hedonism

The four user segments reveal that children's value expression varies primarily in overall expressiveness, with value preference as a secondary distinguishing feature. Together, **User Segment A** and **D** account for approximately 69% of users, suggesting that moderate expression is the norm in this population. Nevertheless, directional tendencies emerge even among low expressors: **User Segment D** leans toward

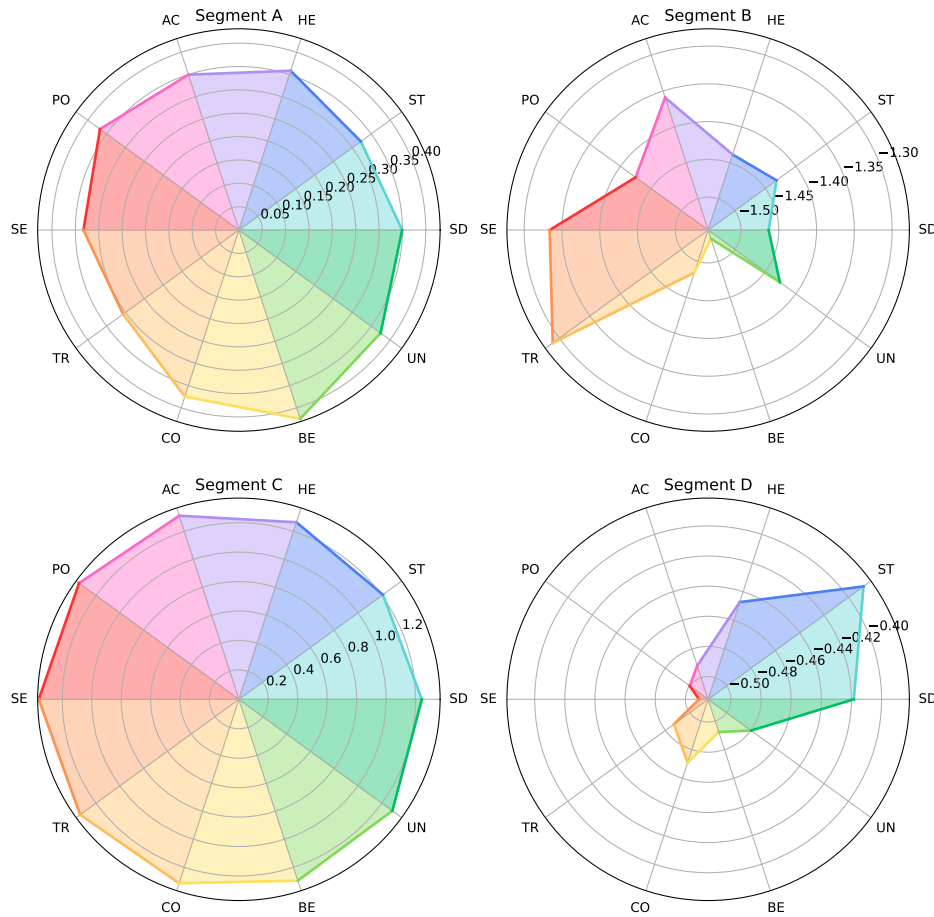


Figure 4.2: Shapes of all 4 user segments.

openness-to-change, while **User Segment B** leans toward conservation. At the high end, **User Segment C** shows that strong value expression tends to be broad rather than selective, with a mild emphasis on order-oriented values. These group-level patterns suggest that children express their values to different extents and motivate a finer-grained analysis to explore internal heterogeneity and outliers.

4.3. Sub-clusters within User Segments

While the four segments capture broad trends in children’s value expression, each group still spans thousands of users. To uncover fine-grained differences within these groups, we apply HDBSCAN and identify sub-clusters using the standardized value profiles. We uncover these meaningful differences within user segments in this section, going beyond the patterns captured by k-means.

4.3.1. User Segment A — Moderate, Balanced Expressor

Within User Segment A, HDBSCAN identifies **19 sub-clusters** alongside a substantial noise group ($n = 1,599$; 49% of cluster members). The high proportion of noise suggests that half of the users within A do not fit into another distinct sub-cluster, but are rather scattered across the space (See Figure 4.3).

The sub-clusters have distinct shapes and value priorities, some with sharp spikes in different directions (e.g A.1 and A.18) and others with a broader coverage of the circumplex (e.g A.11 and A.14). However, among the 19 sub-clusters (each with between 43 and 218 users, see more detailed description of each sub-cluster in Table 4.6), five internal tendencies emerge from the radar profiles (Figure 4.4) and standardized scores (Table 4.5). We indicate similar clusters by coding them using the same color to ease the interpretation and presentation of results:

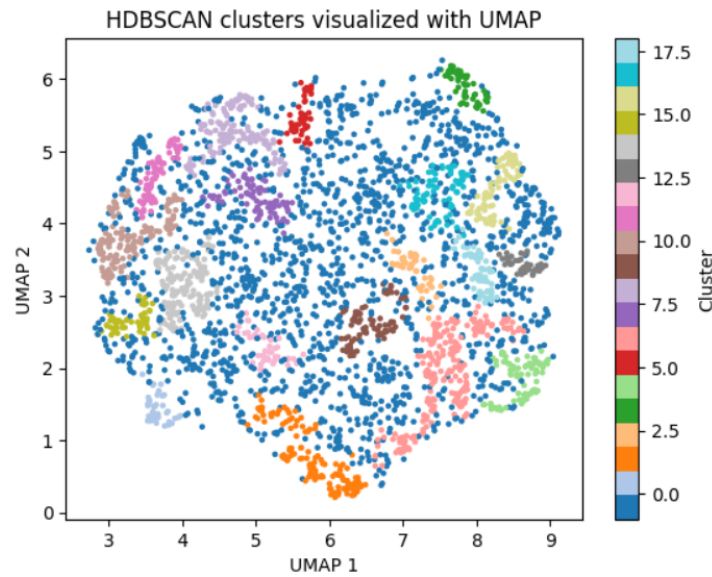


Figure 4.3: Clustered points for User Segment A.

- **Conservation-oriented** (orange; $n = 273users$): Sub-clusters A.0, A.1, and A.4 show spikes on tradition or security relative to all other dimensions, with A.0 reaching $z = 2.182$ for tradition, which is the most extreme score in this segment. These users sit within the moderate segment but express a clear orientation toward social order and stability. However, they still differ in their expressiveness and we see A.4 presenting some interest in social-order values, while for A.0 and A.1, the rest of the values are suppressed.
- **Personal focus with a hint of growth** (purple; $n = 562users$): Sub-clusters A.2, A.3, A.5, A.7, A.8 and A.11 show elevated scores on achievement, hedonism and openness to change values (stimulation, self-direction), while conservation values remain suppressed. This group puts emphasis on personal focus and novelty, with a moderate interest in self-transcendence values. They still differ from one another, for example with A.7 showing strong interest in benevolence, while A.4 ignores this dimension.
- **Personal focus with social suppression** (blue; $n = 266users$): Sub-clusters A.10, A.13, A.16, A.17 and A.18 show high scores on values such as achievement, stimulation and power, which are related to personal focus. However, the degree of coverage across the circumplex differs between them: while A.16 broadly covers self-enhancement, the whole of openness-to-change values and some of self-transcendence, A.18 shows interest in achievement alone, and A.17 more strongly in self-direction.
- **Social focus with openness suppression** (green; $n = 540users$): Sub-clusters A.6, A.9, A.12, A.14 and A.15 show high interest in values that are not within the openness-to-change dimension. These users show elevated scores for security, conformity, tradition, and universalism, while self-direction, stimulation, and hedonism are suppressed. Within an otherwise moderate segment, these subgroups express a clear orientation toward pro-social values and away from seeking novelty and personal interest. While A.14 and A.15 have similar shapes, we see A.9 standing out from the green sub-cluster, with 3 relative elevations (power, stimulation, benevolence) spanning across 3 different higher-order dimensions (self-enhancement, openness-to-change, and self-transcendence respectively). For example, A.12 only peaks in 2 different dimensions, but they are also contrasting (power, universalism).

From these groupings, the orange one stands out as strongly focused on conservation values, with sharp peaks. For the rest, two personal-focused groups emerge: purple and blue. While the purple group is moderately expressive across other values than personal-focused ones, the blue group shows explicit disinterest in social focus, making it a more polarized profile. In particular, blue distinguishes itself from the purple group by having negative scores for conservation and self-transcendence dimensions

Table 4.5: Mean scores for the ten Schwartz value dimensions by sub-cluster, and for the full sample of User Segment A. Values in **bold** indicate scores of at least 2 standard deviations away from the mean of that dimension, whereas *italics* indicate 1 standard deviation difference.

Sub-cluster	SD	ST	HE	AC	PO	SE	TR	CO	BE	UN
A.-1	0.374	0.331	0.356	0.355	0.347	0.319	0.271	0.370	0.428	0.372
A.0	0.627	0.492	0.385	0.525	0.346	0.259	2.182	0.349	0.428	0.674
A.1	0.091	0.045	0.014	0.185	0.025	0.076	1.337	0.055	0.093	0.237
A.2	0.565	-0.176	0.116	0.427	0.386	0.272	-0.098	0.352	0.277	0.187
A.3	0.009	1.078	1.291	-0.295	0.011	-0.146	-0.181	0.036	0.053	-0.150
A.4	-0.013	0.030	-0.143	-0.055	-0.082	0.958	-0.110	-0.062	0.425	-0.030
A.5	0.373	0.518	1.148	0.041	0.160	0.048	0.352	0.269	0.718	0.626
A.6	-0.171	-0.289	-0.182	-0.077	0.315	0.119	0.038	0.431	0.182	0.247
A.7	0.748	0.726	0.469	0.466	0.478	0.522	0.119	0.562	0.818	0.291
A.8	0.545	1.050	1.317	0.295	0.537	0.522	0.397	0.569	0.705	0.524
A.9	-0.018	0.294	0.225	-0.063	0.837	0.384	-0.106	0.474	0.505	0.246
A.10	0.654	0.468	0.377	1.354	0.745	0.630	0.404	0.616	0.662	0.731
A.11	1.289	0.919	0.831	0.910	0.536	0.591	0.517	0.595	0.556	0.775
A.12	0.225	-0.084	0.168	0.146	0.825	0.637	0.347	0.163	0.540	1.118
A.13	0.016	0.473	0.036	0.250	0.325	-0.124	-0.094	-0.123	0.086	-0.117
A.14	0.357	0.410	0.485	0.460	0.966	0.996	0.540	1.055	0.784	0.745
A.15	0.439	0.421	0.654	0.785	0.686	0.561	1.170	0.701	0.741	0.690
A.16	0.275	0.332	0.476	0.260	-0.188	-0.192	-0.281	-0.040	0.059	0.382
A.17	0.918	0.236	0.120	0.001	-0.014	-0.074	-0.134	0.072	0.328	-0.071
A.18	0.038	-0.248	-0.059	1.089	0.104	-0.136	0.054	0.036	-0.035	0.017
Overall	0.349	0.323	0.358	0.350	0.367	0.333	0.306	0.375	0.425	0.375

(Table 4.5). We also see one social-focused group, in green, which appears almost as the inverse of the blue group.

Through these fine-grained differences, we see that within what we consider moderately expressive, balanced children (segment A), distinct sub-groups stand out. By putting 1,599 children in “bucket” A, we overlook all these differences: children who express tradition quite high, such as A.0 and A.1, or children with interest spanning 2 contrasting dimensions, such as A.2, A.12, and A.14. Zooming out, sub-clusters show us that there exist meaningful differences indeed, which go unnoticed when only considering broader user segments. The deeper we look, the more unique value profiles we uncover.

Table 4.6: Description of the 19 sub-clusters identified within User Segment A, grouped by color category. Z-scores reported where values exceed 1 standard deviation from the dimension mean.

Sub-cluster	Group	n	Characteristics
A.-1	Grey	1599	Moderate expression across all dimensions; representative of Segment A overall
A.0	Orange	49	Extreme spike on tradition ($z = 2.18$); suppressed interest across all other dimensions
A.1	Orange	139	Strong interest in tradition ($z = 1.34$); suppressed interest across all other dimensions
A.2	Blue	46	Elevated achievement; suppressed stimulation and tradition
A.3	Purple	60	Strong spikes on stimulation ($z = 1.08$) and hedonism ($z = 1.29$); suppressed achievement and conservation values
A.4	Orange	85	Strong spike on security ($z = 0.96$); strong social-order orientation
A.5	Purple	51	Strong spike on hedonism ($z = 1.15$); moderate expression of remaining values
A.6	Green	218	Suppressed openness-to-change values; relatively elevated conservation and social focus
A.7	Purple	78	Broadly elevated across most dimensions; mild peak on self-direction and stimulation
A.8	Purple	126	Strong spikes on stimulation ($z = 1.05$) and hedonism ($z = 1.32$); broadly elevated across remaining dimensions
A.9	Green	77	Elevated power and conformity; suppressed achievement and tradition
A.10	Blue	134	Strong spike on achievement ($z = 1.35$); broadly elevated across remaining dimensions
A.11	Purple	67	Elevated self-direction ($z = 1.29$); broadly elevated across remaining dimensions
A.12	Green	43	Elevated universalism ($z = 1.12$) and power; suppressed stimulation
A.13	Blue	43	Mildly elevated stimulation and achievement; suppressed conservation and self-transcendence values
A.14	Green	150	Strong spikes on power ($z = 0.97$), security ($z = 1.00$) and conformity ($z = 1.06$); broadly elevated elsewhere
A.15	Green	52	Strong spike on tradition ($z = 1.17$); broadly elevated across remaining dimensions
A.16	Blue	74	Suppressed power, tradition and conservation values; moderate expression elsewhere
A.17	Blue	81	Elevated self-direction; suppressed security, tradition and social focus
A.18	Blue	68	Strong spike on achievement ($z = 1.09$); suppressed stimulation and social focus

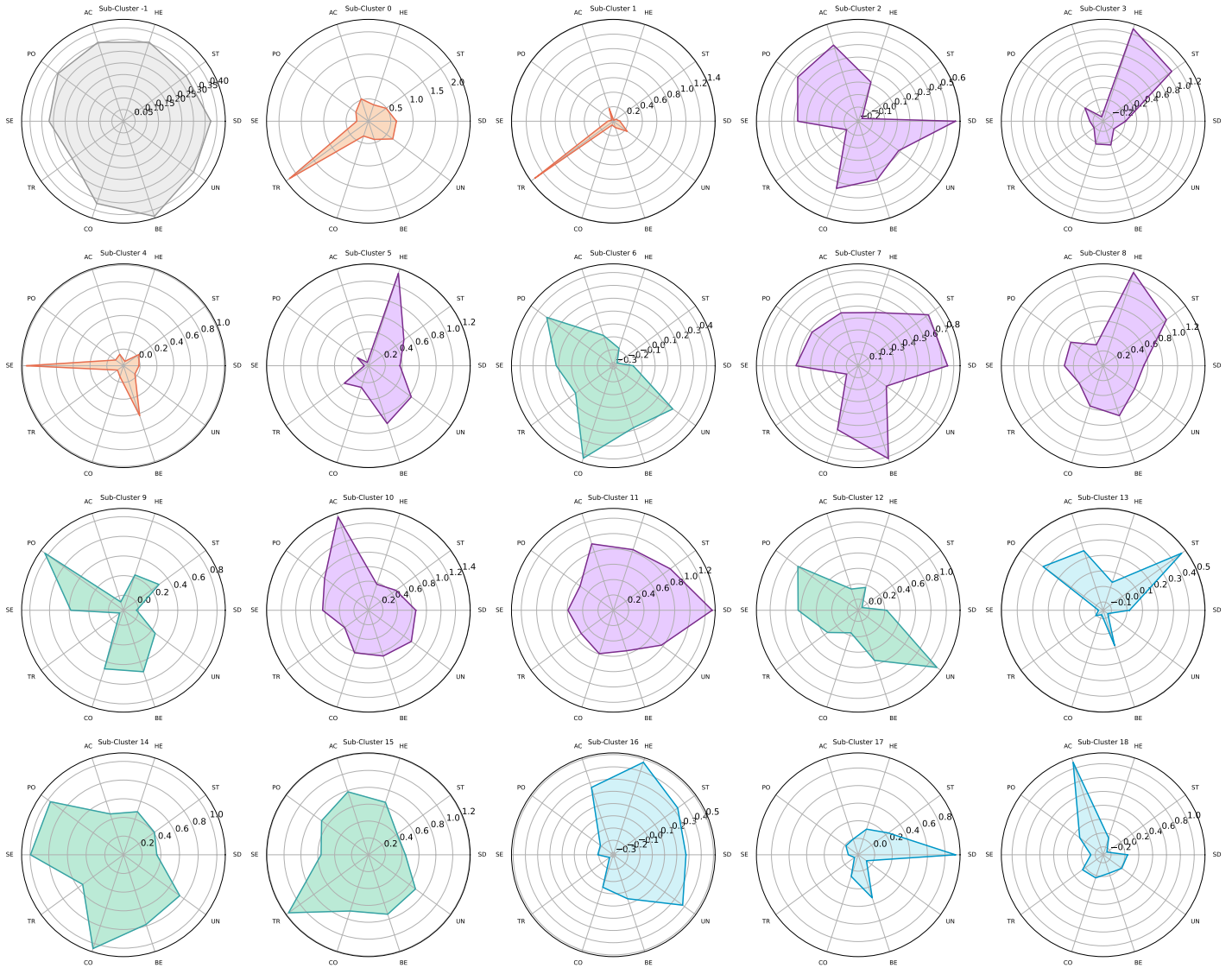


Figure 4.4: Radar plots for sub-clusters within User Segment A.

4.3.2. User Segment B – Disengaged, Conservation-Leaning Expressor

Within User Segment B, HDBSCAN identifies **5 sub-clusters**, with a noise proportion of $n = 294$ (23.5% of cluster members). The lower noise rate suggests that more users within B fall into more distinguishable sub-clusters than A.

The sub-cluster profiles (Figure 4.5 and Table 4.7) reveal differentiation primarily in which values are least suppressed within an overall low-expression context. All sub-clusters remain well below the population mean, just as the overall segment, but they differ in both intensity and direction. This is consistent with the segment-level finding that conservation values are the relative peak for this group, but sub-clusters show that individual users vary in how that lean manifests.

The largest sub-cluster, **B.2** ($n = 385$), is the least suppressed group overall, with scores ranging from -0.95 (*universalism*) to -1.30 (*stimulation*). Within this low-expression context, *universalism*, *security*, and *power* are the least suppressed, suggesting a mild relative lean toward self-protection and broad concern for others. This resembles the conservation lean characteristic of segment B, putting no emphasis on stimulation, the opposing value to tradition. Similarly, **B.4** ($n = 127$) shows clear conservation lean, while suppressing the openness-to-change dimension the most (*self-direction*: -1.66, *stimulation*: -1.60, while *security*: -1.27, *tradition*: -1.35).

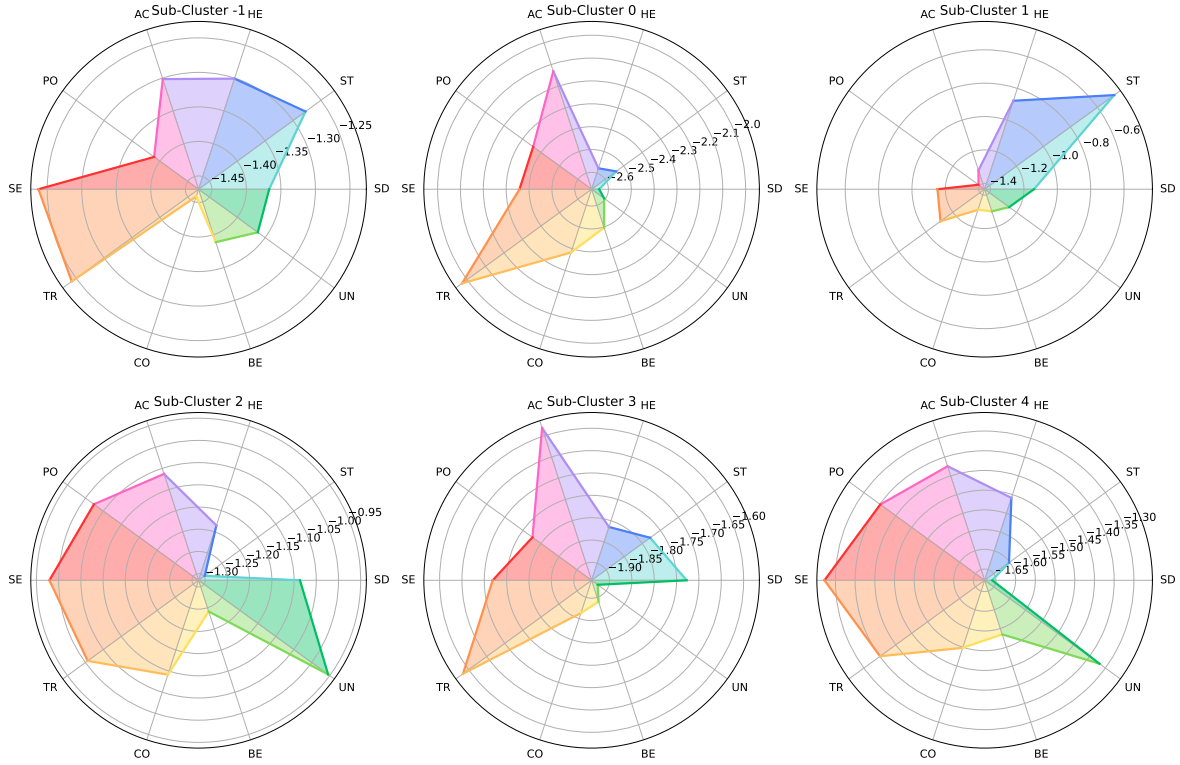


Figure 4.5: Radar plots for sub-clusters within User Segment B.

Table 4.7: Scores for the groups identified within User Segment B (mean, standard deviation).

Cluster	SD	ST	HE	AC	PO	SE	TR	CO	BE	UN
B.-1	-1.37 ± 0.47	-1.28 ± 0.51	-1.3 ± 0.48	-1.3 ± 0.55	-1.39 ± 0.42	-1.24 ± 0.44	-1.24 ± 0.41	-1.46 ± 0.46	-1.39 ± 0.55	-1.36 ± 0.45
B.0	-2.64 ± 0.61	-2.54 ± 0.55	-2.58 ± 0.57	-2.13 ± 0.55	-2.36 ± 0.56	-2.36 ± 0.52	-1.97 ± 0.5	-2.38 ± 0.45	-2.5 ± 0.58	-2.61 ± 0.47
B.1	-1.14 ± 0.36	-0.48 ± 0.52	-0.88 ± 0.48	-1.32 ± 0.28	-1.39 ± 0.28	-1.15 ± 0.32	-1.11 ± 0.31	-1.31 ± 0.32	-1.3 ± 0.33	-1.25 ± 0.31
B.2	-1.09 ± 0.45	-1.3 ± 0.35	-1.19 ± 0.31	-1.06 ± 0.4	-1.02 ± 0.33	-0.98 ± 0.35	-1.01 ± 0.39	-1.09 ± 0.39	-1.24 ± 0.35	-0.95 ± 0.38
B.3	-1.73 ± 0.56	-1.78 ± 0.44	-1.81 ± 0.44	-1.58 ± 0.54	-1.78 ± 0.33	-1.72 ± 0.37	-1.58 ± 0.43	-1.86 ± 0.41	-1.89 ± 0.43	-1.92 ± 0.4
B.4	-1.66 ± 0.34	-1.6 ± 0.33	-1.46 ± 0.3	-1.37 ± 0.36	-1.35 ± 0.36	-1.27 ± 0.32	-1.35 ± 0.26	-1.5 ± 0.37	-1.53 ± 0.32	-1.32 ± 0.36

B.3 ($n = 195$) presents a relative peak for *achievement* in the radar plot, but it is actually below the mean score for this user segment. We see that although this group suppresses values more than the average within the user segment, it resembles the overall shape of **B**, apart from the relative peak on *achievement*. By suppressing self-transcendence and conformity values most, (*universalism*: -1.92, *benevolence*: -1.89, *conformity*: -1.86), while *tradition* (-1.58) and *achievement* (-1.58) this sub-group distances itself from prosocial and socially compliant values. The smallest sub-cluster, namely **B.0** ($n = 118$), has a similar shape and represents the most disengaged users, with scores ranging from -1.97 (*tradition*) to -2.64 (*self-direction*).

One group is clearly distinctive from the others, namely **B.1** ($n = 129$). In this case, its shape does not resemble segment B's due to the elevation in *stimulation*. This dimension is far less suppressed (-0.48), while the remaining values cluster tightly around -1.1 to -1.4. This isolated spike toward stimulation within an otherwise disengaged profile stands out and suggests that a subset of Segment B's users does express interest for novelty and excitement.

4.3.3. User Segment C – Engaged, Prosocial Expressor

Within User Segment C, we identified **4 sub-clusters**, with a low noise proportion ($n = 270$; 15.97% of cluster members). This low noise rate and small number of sub-clusters suggest that Segment C is the most internally homogeneous group. Users who express values broadly and intensely tend to do so in similar ways, leaving little room for distinct subgroups to emerge.

Table 4.8: Scores for the groups identified within User Segment C (mean, standard deviation).

Cluster	SD	ST	HE	AC	PO	SE	TR	CO	BE	UN
C.-1	1.09 ± 0.55	1.15 ± 0.6	1.26 ± 0.6	1.24 ± 0.52	1.4 ± 0.61	1.23 ± 0.53	1.53 ± 0.82	1.26 ± 0.51	1.32 ± 0.47	1.36 ± 0.54
C.0	1.01 ± 0.54	0.93 ± 0.57	0.97 ± 0.55	1.04 ± 0.57	1.04 ± 0.5	1.04 ± 0.53	0.91 ± 0.59	1.02 ± 0.51	1.05 ± 0.43	0.99 ± 0.49
C.1	1.8 ± 0.52	1.76 ± 0.58	1.88 ± 0.65	1.87 ± 0.55	1.92 ± 0.53	2.18 ± 0.64	2.15 ± 0.63	1.93 ± 0.53	1.78 ± 0.38	1.93 ± 0.53
C.2	1.14 ± 0.43	0.88 ± 0.37	1.12 ± 0.33	1.38 ± 0.39	1.24 ± 0.43	1.06 ± 0.34	1.31 ± 0.29	1.67 ± 0.45	1.32 ± 0.32	1.43 ± 0.51
C.3	1.43 ± 0.5	1.47 ± 0.55	1.39 ± 0.52	1.48 ± 0.5	1.49 ± 0.49	1.5 ± 0.48	1.39 ± 0.51	1.42 ± 0.44	1.39 ± 0.42	1.31 ± 0.47

As visible in Figure 4.6 and Table 4.8, scales differ slightly across sub-clusters, but most profiles remain full and expansive, closely resembling segment C. The largest sub-cluster, **C.0** ($n = 754$; 44.6%), represents the core of User Segment C with scores hovering around 1.0 across all dimensions. No value stands out dramatically, with tradition (0.91) and stimulation (0.93) marginally lower than the rest. This group embodies the broad, undifferentiated high expression characteristic of the user segment. **C.3** ($n = 311$) follows this pattern too, showing a uniformly elevated and balanced profile, with scores ranging from 1.31 (*universalism*) to 1.50 (*security*). It can be considered an amplified version of **C.0**, expressing all values slightly more intensely.

C.1 ($n = 303$) represents the most intensely expressive sub-cluster in the entire dataset, with scores well above the already-high segment average across all dimensions. *Security* (2.18) and *tradition* (2.15) are the most prominent, reinforcing the order-oriented tendency identified at the segment level. This sub-cluster represents children whose book reviews are much richer in expressiveness across all dimensions.

C.2 ($n = 52$) is the smallest sub-cluster and the only one within User Segment C with a meaningfully distinct shape. While remaining highly expressive overall, it stands out for suppressing *stimulation* (0.88) relative to all other dimensions, while showing elevated *conformity* (1.67), *universalism* (1.43), and *achievement* (1.38). This pattern suggests a subset of highly expressive users who emphasize prosocial and rule-following values while showing comparatively little interest in novelty and excitement.

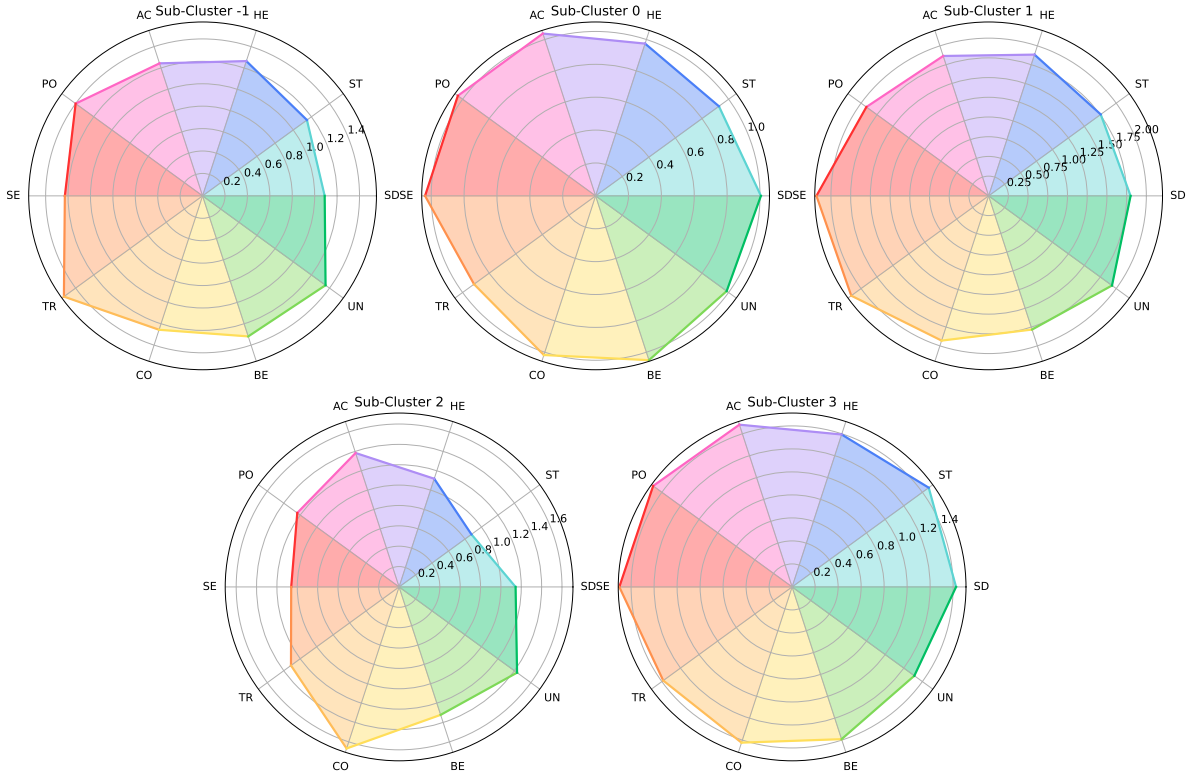
**Figure 4.6:** Radar plots for sub-clusters within Segment C.

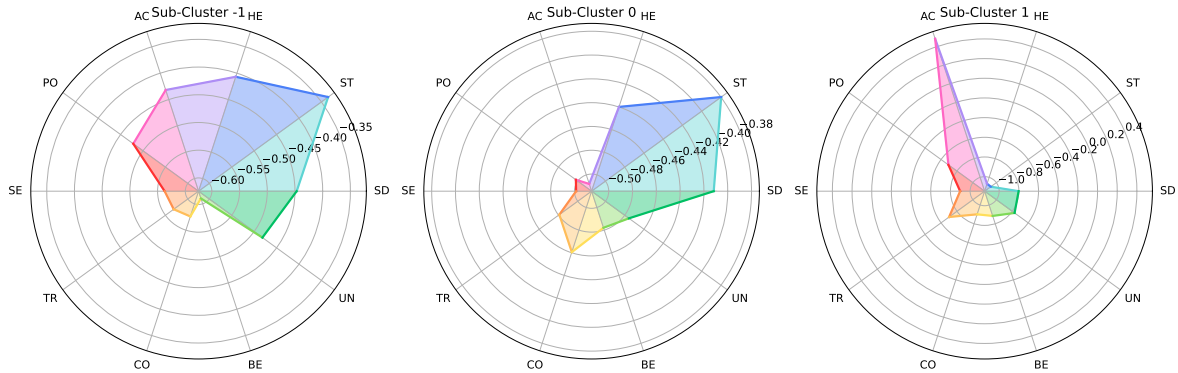
Table 4.9: Scores for the groups identified within User Segment D (mean, standard deviation).

	SD	ST	HE	AC	PO	SE	TR	CO	BE	UN
Cluster										
D.-1	-0.45 ± 0.57	-0.33 ± 0.73	-0.41 ± 0.58	-0.43 ± 0.61	-0.48 ± 0.5	-0.56 ± 0.42	-0.57 ± 0.46	-0.58 ± 0.42	-0.61 ± 0.4	-0.48 ± 0.44
D.0	-0.41 ± 0.56	-0.38 ± 0.6	-0.44 ± 0.51	-0.51 ± 0.53	-0.5 ± 0.47	-0.5 ± 0.47	-0.48 ± 0.5	-0.46 ± 0.49	-0.48 ± 0.48	-0.48 ± 0.49
D.1	-0.8 ± 0.34	-1.07 ± 0.33	-1.07 ± 0.25	0.48 ± 0.41	-0.69 ± 0.44	-0.9 ± 0.32	-0.7 ± 0.35	-0.9 ± 0.41	-0.88 ± 0.38	-0.77 ± 0.36

4.3.4. User Segment D — Selective, Openness-to-Change Expressor

Within User Segment D, HDBSCAN identified **2 sub-clusters**, with negligible noise ($n = 123$, 3.66% of cluster members). The extremely low noise proportion and minimal sub-cluster count make this the most internally homogeneous segment. The dominant sub-cluster, **D.0** ($n = 3,193$), displays the expected pattern of below-average expression with a relative peak on openness-to-change, with scores ranging from -0.38 for *stimulation* to -0.56 for *security*). The smallest group, D.1 ($n = 46$ members), deviates from this by showing a strong interest in *achievement*, while suppressing all other dimensions (in Figure 4.7 we see achievement is the only value with a positive z-score). This represents a subset of users who do not fit the openness-to-change tendency of the segment, but are rather ambitious, distinguishing themselves by valuing achievement way more than their peers.

Sub-cluster analysis reveals that the four segments vary considerably in their internal heterogeneity. **User Segment A** is the most fragmented, with 19 sub-clusters reflecting the wide variety of value configurations that can produce a moderate, balanced profile. **User Segment C** is the most cohesive in terms of shape, given that its sub-clusters differ in intensity but not direction. **User Segment D** is the most homogeneous in terms of distribution, with 95% of users concentrated in a single sub-cluster. **User Segment B** occupies a middle ground, with five distinguishable patterns of suppression. Together, these results suggest that children’s value expression is structured at multiple levels of granularity, from broad expressiveness tendencies down to specific value preferences that form meaningful subgroups.

**Figure 4.7:** Radar plots for sub-clusters within Segment D.

4.4. Evaluating Value Alignment of Children and Recommendations

Moving on to the results of the recommender experiment described in Section 3.3, we investigate how different algorithms reflect children’s values.

4.4.1. Value coverage

To explore whether the full spectrum of values is represented in recommendations, we report **coverage** and **skew**, as defined in Section 3.3. Results are reported in Table 4.10.

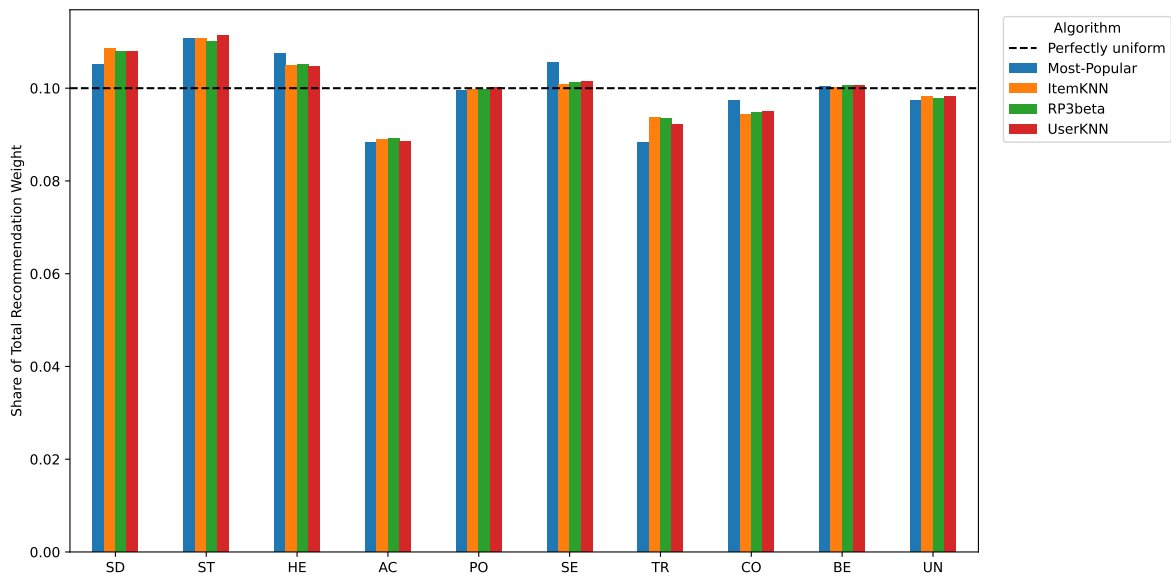
All four algorithms receive entropy scores above 0.99, confirming that all values are covered almost uniformly and that none of the algorithms concentrate on a small subset of values. To examine differences at the level of individual value categories instead, we look at the share that each value gets out of the total recommendation weight, plotted in Figure 4.8. Here we observe that values like power, security and benevolence are very close to being covered perfectly uniformly by all algorithms. Self-direction,

Table 4.10: Value coverage metrics per algorithm.

Algorithm	Normalized entropy	Skew
MostPop	0.999049	1.24
ItemKNN	0.999162	1.25
UserKNN	0.999116	1.27
RP3beta	0.998878	1.25

stimulation and hedonism stand out as being covered more, at the expense of achievement, tradition, conformity and universalism. This indicates that on a population level, values are similarly prominent, with minor and consistent deviations toward certain values regardless of which algorithm is used.

In terms of skew (calculated as $\frac{max}{min}$), which shows the degree to which values are balanced in the recommendations, UserKNN shows the highest ratio (ratio = 1.27), while MostPop shows the most even spread (ratio = 1.24). This indicates that although there are some differences between recommendations based on algorithms, the degree of emphasis does not vary considerably when it comes to value coverage.

**Figure 4.8:** Value dimension coverage per recommender algorithm.

4.4.2. Amplification and Suppression

To explore whether some value dimensions are over- or under-represented in recommendations relative to what children express, we compute the log-ratio between recommendations and children's expressed value prominence (normalized value scores per value dimension). Across all four algorithms, security is the most consistently amplified value (change of 0.08 to 0.14, roughly 6–10% relative over-representation), while tradition was the most consistently suppressed (-0.02 to -0.08). This is visualized in the heatmap of Figure 4.9, where the most pronounced values are security and tradition. Since this effect is independent of algorithm choice, it is unlikely to be an artifact of any single algorithm's design, and may instead reflect a systematic mis-alignment between children's expressed value profiles and how recommendations are generated.

Figure 4.10 shows that Most-Popular produces the largest amplification of any algorithm-value pair (security), while UserKNN produces the largest suppression (tradition). ItemKNN and RP3beta show similar amplification and suppression patterns, with these typically being modest. Security and power are amplified across all four algorithms, while tradition, conformity, universalism, and self-direction are consistently suppressed. Notably, these groups of values are not scattered across the Schwartz circumplex, but are next to each other. This suggests the algorithms do not amplify or suppress conflicting values, but rather favor one value dimension over others.

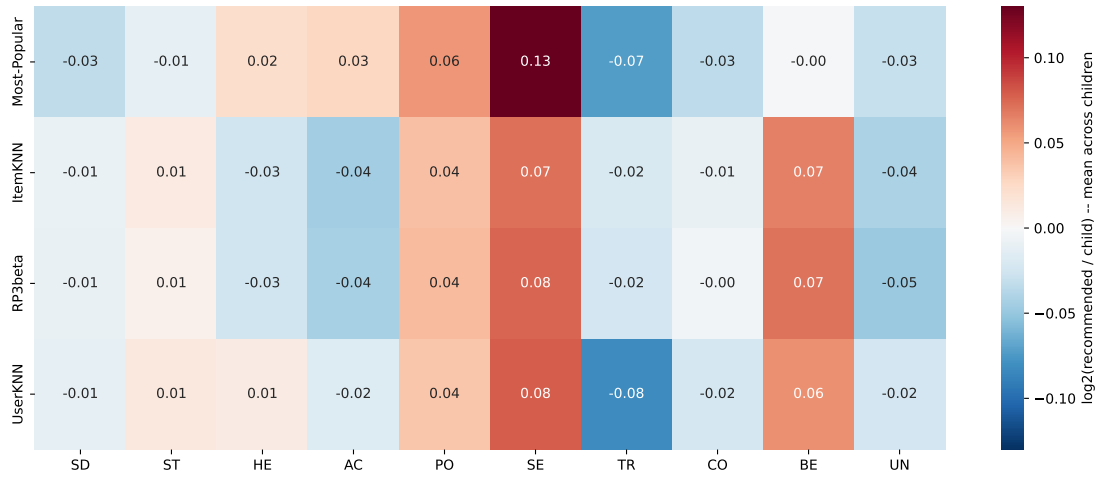


Figure 4.9: Heatmap of value amplification and suppression per algorithm.

Looking at algorithms individually adds further nuance to this pattern. Most-Popular favors security, power, hedonism, and achievement—all adjacent values in Schwartz’s circumplex, largely covering the self-enhancement dimension and self-protection. This is also the only algorithm amplifying achievement. ItemKNN, UserKNN, and RP3beta instead seem more social-focused, amplifying benevolence over power. While the latter three are more similar, what stands out is that hedonism is amplified for UserKNN and tradition is more strongly suppressed by UserKNN. Thus, different algorithms may be perceived differently by users, although they seem to be covering all values pretty uniformly. Interestingly, self-direction and stimulation appear well-represented in the catalog coverage analysis (Figure 4.8), yet both are suppressed in the recommendations for children.

In terms of distortion (Table 4.11), which highlights how much recommended items’ value emphasis deviates from each child’s expressed value profile, Most-Popular achieves the lowest average distortion (0.2094), indicating its recommendations are, on average, the closest match to individual children’s expressed values. UserKNN follows (0.2312), while RP3beta and ItemKNN show the highest and nearly identical distortion (0.2493 and 0.2497, respectively). A Friedman test confirmed a statistically significant difference in distortion across algorithms ($p < 0.001$), though the effect size was small (Kendall’s $W = 0.058$).

Table 4.11: Mean per-child value distortion by algorithm.

Algorithm	Mean Distortion
Most-Popular	0.2094
UserKNN	0.2312
RP3beta	0.2493
ItemKNN	0.2497

Furthermore, in Figure 4.11 we see the same trend across algorithms: a child whose top value is tradition or achievement will find it pushed to the bottom of their recommendations’ ranking, reinforcing our earlier finding that tradition is the most suppressed value. Generally, for the openness-to-change dimension, the child’s and recommendations’ top priorities align, while conservation sits at the bottom. Once again, we see that the different algorithms reflect the top priority similarly, with a slight difference for benevolence: most algorithms rank it in the middle while Most-Popular ranks it lower.

While Most-Popular produced the single largest amplification (security) and one of the largest suppressions (tradition) of any algorithm-value pair, it still achieves the lowest overall distortion. This is consistent with ItemKNN’s amplification profile (Figure 4.9), which suppresses a wider set of categories at moderate magnitude rather than concentrating suppression on a single outlier value, potentially explaining why it attains the highest distortion despite having no single amplification or suppression

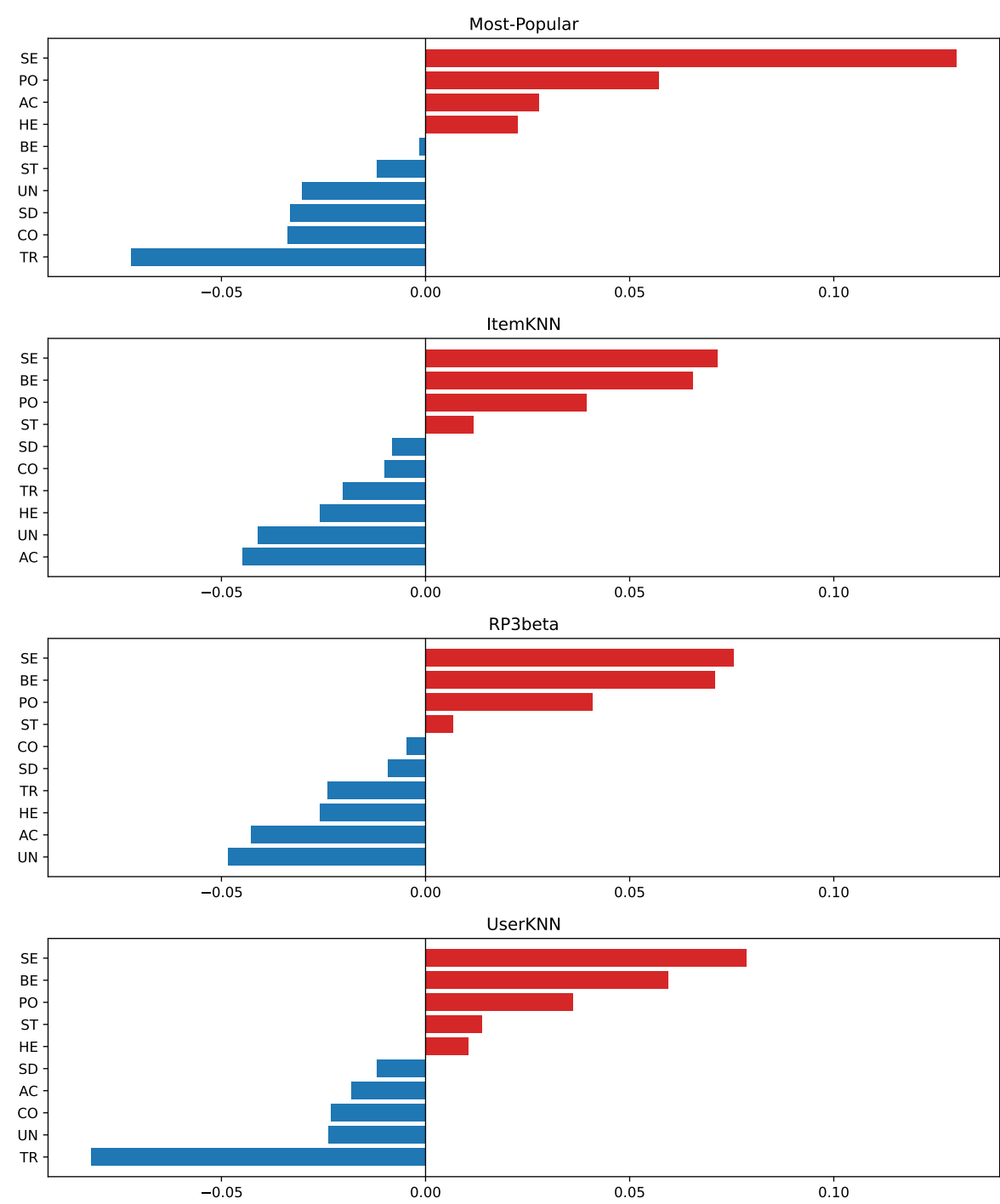


Figure 4.10: Bar graph of value amplification and suppression per algorithm.

bar as extreme as Most-Popular’s.

On User-Segment Level

Breaking this down by user segment (Figure 4.12) shows that security and power remain the most amplified values across every segment, underscoring that these population-level effects reflect on different user segments. For User Segment B, power is amplified even more than security, reflecting a conservation-leaning segment where security is already relatively high. Tradition is no longer suppressed across all algorithms, with ItemKNN amplifying it for clusters B and D. The value that gets

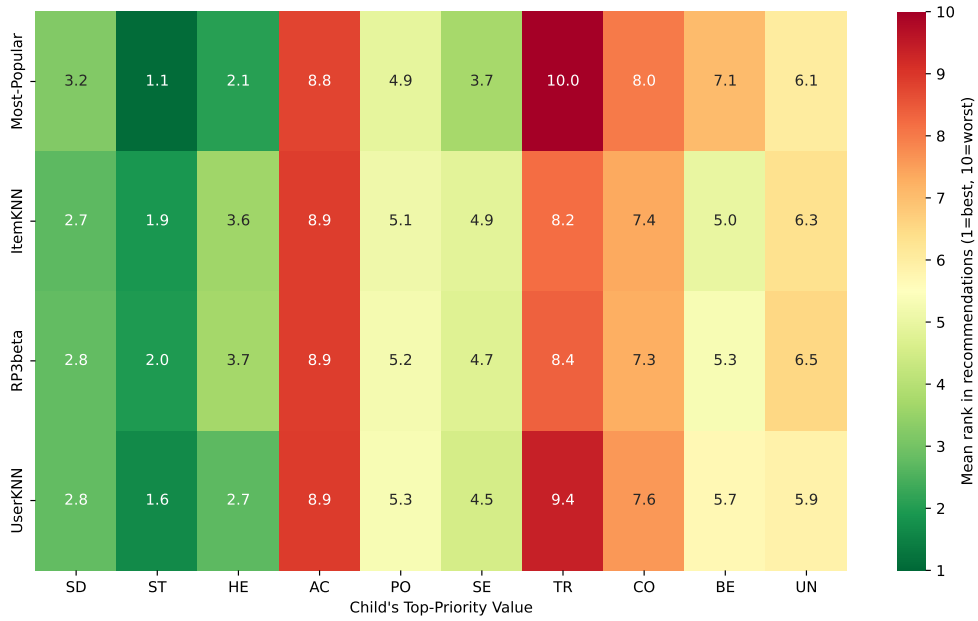


Figure 4.11: Heatmap showing where a child's top value ends up when the values of their recommended items.

most suppressed at the segment level is achievement for segment D, especially by UserKNN, although RP3beta is suppressing achievement more than UserKNN when looking at the population as a whole (Figure 4.9).

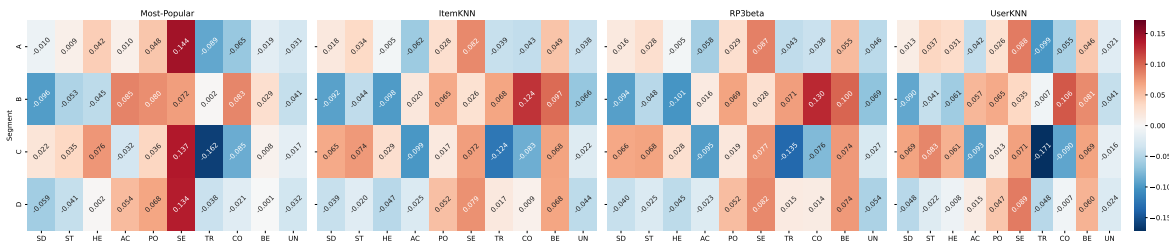


Figure 4.12: Heatmaps of value amplification and suppression per algorithm, per user segment

Regarding per-child distortion, reported in Table 4.12, Most-Popular achieves the lowest distortion in segments A and D, while for segments B and C UserKNN distorts the least. Notably, neither ItemKNN nor RP3beta is ever the best-scoring algorithm in any individual segment, consistent with them being the two most-distorting algorithms at the population level, where their overall means are nearly identical (0.2493 vs. 0.2497). However, in the case of segment C, RP3beta performs similarly to the best one (UserKNN), while ItemKNN remains the one with highest distortion. This shows that the overall population mean (Table 4.11) can mask meaningful differences between algorithms once individual segments are considered; two algorithms with similar overall means are not necessarily performing similarly for any specific group of children. Taken together, this segment-level breakdown reinforces that no single algorithm is uniformly best at preserving values across all children. The strategy that aligns most with children's values depends on which segment is considered, although Most-Popular and UserKNN outperform ItemKNN and RP3beta for each segment.

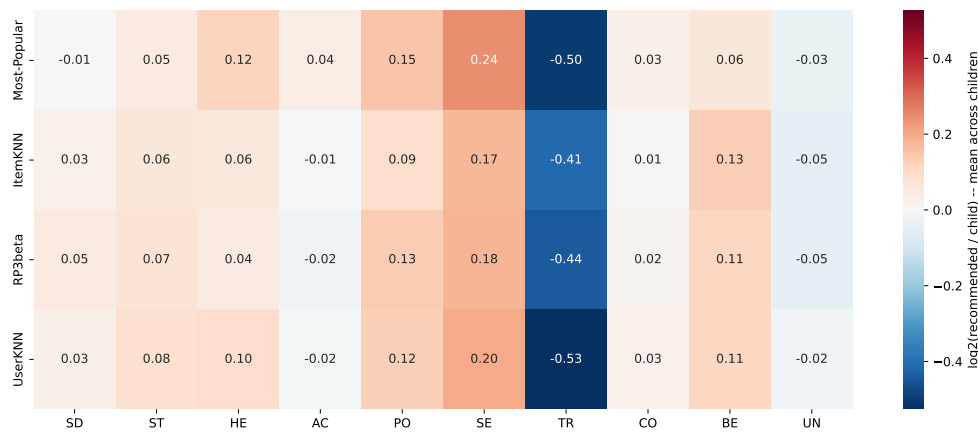
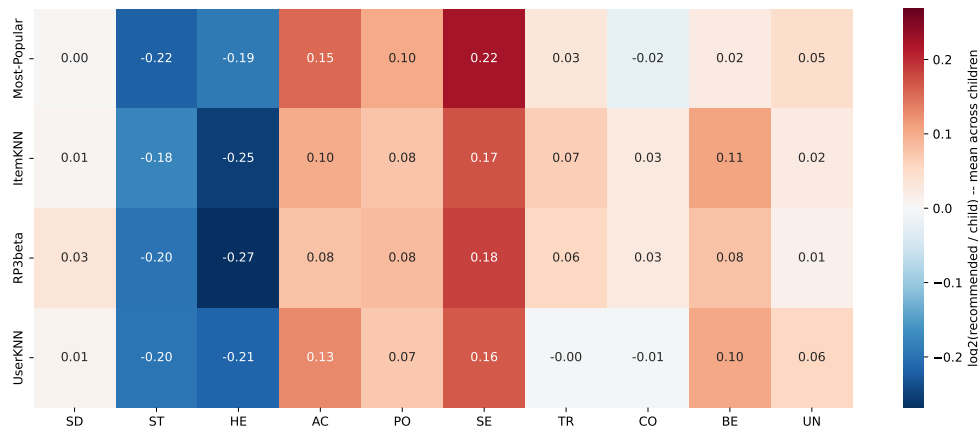
On Sub-cluster Level

Zooming in from the segments to sub-clusters reveals that this distortion pattern is not uniform, but rather intensifies or shifts depending on the group's value composition. For sub-clusters, we see that A.0 (Figure 4.13) exhibits a particularly pronounced version of the population-level pattern, with security consistently amplified (average log-ratio between 0.17 and 0.24) and tradition strongly suppressed (average log-ratio between -0.53 and -0.41) across all four algorithms. In contrast, sub-cluster A.3 (Figure

Table 4.12: Mean per-child value distortion by algorithm and user segment.

Segment	Most-Popular	ItemKNN	RP3beta	UserKNN
A	0.1555	0.2042	0.2303	0.1743
B	0.2270	0.2208	0.2455	0.2041
C	0.1902	0.2643	0.1773	0.1694
D	0.1463	0.1926	0.1875	0.1719

4.14) shows stimulation and hedonism consistently suppressed (log-ratio between -0.27 and -0.18), while achievement, power, and security are amplified (log-ratio between 0.07 and 0.22). Notably, these patterns do not appear to show adaptation to the value profiles of the respective sub-clusters. The algorithms produce broadly similar amplification and suppression patterns across the two groups. This suggests that the observed value distortions are driven more by systematic properties of the recommendation process than by adaptation to value differences between user sub-clusters.

**Figure 4.13:** Heatmaps of value amplification and suppression per algorithm, for sub-cluster A.0.**Figure 4.14:** Heatmaps of value amplification and suppression per algorithm, for sub-cluster A.3.

Some sub-clusters exhibit deviations in values that are part of the same higher-order dimension. Sub-cluster D.1 (Figure 4.15), for instance, shows strong suppression of achievement across all algorithms (log-ratio between -0.36 and -0.33), alongside amplification for power (log-ratio between 0.02 and 0.10), which together constitute the self-enhancement dimension. This provides an example where the distortion is not simply a uniform shift across a higher-order dimension, but instead affects adjacent values differently. Overall, these examples indicate that value amplification and suppression are not

uniform across users: their magnitude and configuration can vary substantially between sub-clusters, while often remaining consistent across recommendation algorithms.

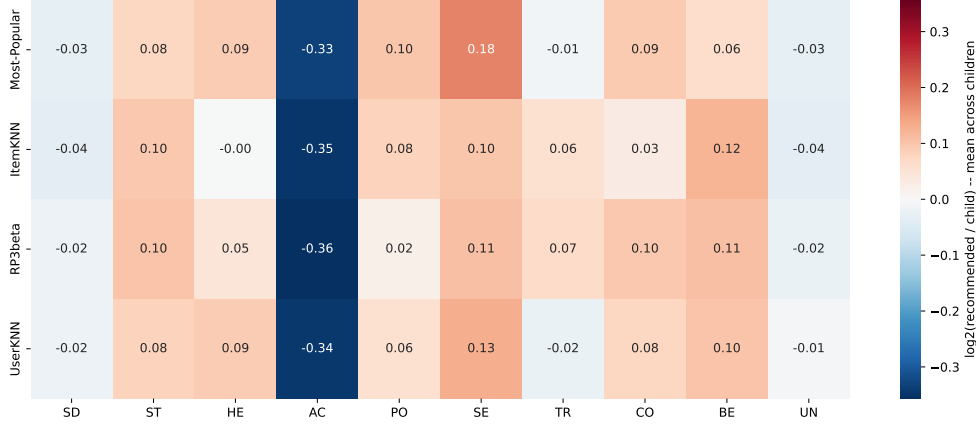


Figure 4.15: Heatmaps of value amplification and suppression per algorithm, for sub-cluster D.1.

4.4.3. Accuracy

Accuracy is low for all algorithms. Mean $nDCG@10$ takes the values of 0.020 (Most-Popular), 0.044 (ItemKNN), 0.050 (UserKNN), and 0.046 (RP3beta) for each algorithm specifically. Given the sparsity of the interaction data, for Most-Popular only 7.66% of users had at least one hit in top-10, only 13.70% for ItemKNN, 14.06% for RP3beta and 15.83% for UserKNN, which explains the low scores for $nDCG$.

Table 4.13: $nDCG@10$ for each user segment and recommendation algorithm.

User Segment	Most-Popular	ItemKNN	RP3beta	UserKNN
A	0.0216	0.0232	0.0117	0.0185
B	0.0000	0.0170	0.0702	0.0651
C	0.0215	0.0415	0.0411	0.0377
D	0.0412	0.0379	0.0449	0.0794

Between segments, we see in Table 4.13 that ItemKNN performs best for segment A and C, RP3beta performs considerably better for B, and UserKNN performs better for segment D. To test whether value-aligned recommendations come at a cost to accuracy, we compute Spearman rank correlations between $nDCG$ and both amplification and distortion, per algorithm and per user segment (see Table 4.14). At the population level, no algorithm showed a meaningful relationship between accuracy and amplification ($|r| \leq 0.0191$, $p > 0.06$) or distortion ($|r| \leq 0.0166$, $p > 0.1048$). Repeating this analysis within each user segment (16 tests per metric) yielded no correlations surviving a Bonferroni-adjusted significance threshold, with all effect sizes remaining negligible ($|r| < 0.05$). These results indicate that, in this dataset, improving value-alignment does not come at a measurable cost to recommendation accuracy, nor does it occur as a side effect of accuracy optimization.

Overall, all algorithms cover the full value space, although they differ in how concentrated their recommendations are within it, with Most-Popular exhibiting the greatest skew. Across all algorithms, security is systematically over-represented while tradition is under-represented. In terms of recommendation accuracy, UserKNN achieves the highest $nDCG@10$, although absolute accuracy remains relatively low across all algorithms. Finally, we find no evidence of a trade-off between recommendation accuracy and value alignment, indicating that improving accuracy does not necessarily come at the expense of alignment with the considered values.

Table 4.14: Spearman correlations between the evaluated metrics and the corresponding variables.

Algorithm	Metric	Spearman's r	Spearman's p
Most-Popular	amplification	-0.0191	0.0619
Most-Popular	distortion	-0.0166	0.1048
ItemKNN	amplification	-0.0139	0.1746
ItemKNN	distortion	-0.0150	0.1437
RP3beta	amplification	-0.0023	0.8204
RP3beta	distortion	-0.0014	0.8944
UserKNN	amplification	-0.0001	0.9896
UserKNN	distortion	0.0006	0.9559

DISCUSSION

In this chapter, we interpret the results presented in Chapter 4 in light of the two research questions that guide our research, discuss their implications for children and for the design of recommender systems, and reflect on the limitations and opportunities for future work. While we focus on books here, the insights can be further applied to other domains, such as movies, social media or news.

RQ1: What values do children prioritize, as expressed through their book reviews?

General trends Children’s reviews do not reveal that a single value is the most important for them. Instead, they reflect a relatively balanced mix of values, as we see that mean scores across all ten Schwartz dimensions fall within a narrow segment (between 0.2538 and 0.3429). Stimulation stands out only marginally as the highest mean, and hedonism has the maximum individual value score. This suggests children’s value expression tends to prioritize openness to change values, which is in line with longitudinal studies done with children [18, 21]. These studies suggest that children tend to emphasize self-transcendence and openness to change values (especially stimulation), while ascribing less importance to conservation values. Self-transcendence values are not as prevalent in our data, and we observe that tradition registers the widest spread and highest standard deviation across dimensions.

Given the high standard deviation, tradition is found to be the most polarizing value in our study, meaning some children’s reviews engage with it strongly, while others barely at all. Two readings of this polarization are worth distinguishing. The first, more intuitive reading is that tradition constitutes a genuine axis of disagreement among children, plausibly reflecting the diversity of cultural, religious, and family backgrounds in children [2]. The second, more cautious reading draws on prior validation work with value instruments for children, which finds that even when children can make consistent value choices using picture-based formats, their responses can deviate from Schwartz’s circular ordering, specifically around the placement of tradition and conformity in the self-transcendence region rather than conservation [19]. This is considered a small deviation that may appear in young children, as these value items express motivations similar to benevolence for them [18]. Thus, if children’s understanding of “tradition” does not map as cleanly onto a single coherent construct as it does for adults, part of the polarization observed here could reflect instability in how the value is understood by children and expressed in their language, rather than differences in priorities alone. Since our method scores tradition against a single definition, and children comprehend this value differently, it is possible that some younger children use language that is closer to the benevolence definition when they express tradition-related sentiments and the other way around, widening the spread for tradition.

At first glance, one might easily rush into concluding children’s values are broadly homogeneous, with slight differences in emphasis. As discussed above, looking at our data based on a simple population mean confirms the trend that has been identified through psychological studies. However, our further analysis shows that we are averaging over meaningfully different groups and putting all children in one homogeneous box that does not reflect lived realities.

User Segments Clustering children’s value profiles yields four significantly distinct and substantially separated segments, suggesting that they capture meaningful differences in children’s value orientations. We see these differences mainly along two axes: an expressiveness axis (the degree to which they

express values generally) and a value-preference axis (which higher-order value dimension they lean towards). Within this structure, the two below-average-expressiveness user segments (disengaged) lean in opposite directions, one toward conservation (Segment B) and one toward openness-to-change (Segment D), while the one broadly and intensely expressive group (Segment C) shows a mild lean toward order and self-protection rather than expressing all values with equal intensity.

Sub-clusters Sub-cluster analysis within segments adds a further layer: internal heterogeneity is itself unevenly distributed, with the moderate, balanced segment fragmenting into 19 distinguishable sub-clusters spanning nearly the full space of conservation, openness-to-change, and personal-focused orientations, while the most expressive segment remains relatively cohesive, differing mainly in intensity rather than direction. In other words, a segment's apparent "balance" can itself be an average over children pulling in many different directions. This shows that considering children as a group conceals individual differences, and that even the more refined step of considering them as segments can still flatten real differences unless we look at how uniform each segment is.

Contextualization To see how our method identifies values, we take an example of what a review and value expression looks like. We contextualize this with the more conventional method of the dictionary approach. The dictionary-based approach uses frequency count as an indication for value importance [60], and allows us to point to words that are typically associated with certain values, thus providing insights into what our method picks up on.

In the review from below, the dictionary flags eleven word tokens across six of the ten Schwartz values (which we color-coordinate following the colors from 2.1), with power, achievement, security, and conformity receiving no hits at all, since none of their associated words occur verbatim in the text. By construction, the dictionary approach can only flag a value where its predefined vocabulary is matched exactly, so its output is essentially a frequency count over these six categories, with self-direction (4 hits) and benevolence (3 hits) appearing most salient.

The embedding-based value profile, as leveraged in our study, tells a partly different story. Stimulation is identified as the most prominent value (with a score of 0.337), even though the dictionary approach identified only a single stimulation-related word (*courage*). This may be because the review is saturated with adventure-related content such as the shark attack, "intense", "suspenseful", "scary", the tribal-prophecy subplot and the ghost-hunting storyline. None of these tokens are in the dictionary's stimulation list, but collectively they build a semantic context that an embedding model seems to pick up on. With stimulation being defined as looking for excitement and challenge in life, our approach seems to get more nuances, rather than explicit trait words, which the dictionary is limited to.

A similar pattern may explain the comparatively high security score (0.279) despite zero dictionary hits. Security in Schwartz's framework relates to safety, stability, and harmony. The review contains small routines that denote family stability (reading with the family "every night at dinner"). These are thematic cues, so a dictionary approach misses them by design, while the embedding representation seems to capture this meaning.

The reverse discrepancy is also present: the dictionary approach would point you towards *benevolence* being one of the most prominent values (three hits: *family*, *love*, *mom*), yet in the embedding profile benevolence is the second-lowest score (0.199). A likely explanation is that dictionary matching is context-blind, registering the token *love* regardless of its object, so "i love this book" counts the same as expressing love for a person, even though Schwartz's benevolence construct is specifically about preserving and enhancing the welfare of those close to you, not affective enthusiasm toward objects or activities. Likewise, *mom* and *family* are treated as evidence of benevolence purely by co-occurrence, irrespective of whether the sentence actually expresses care, warmth, or prosocial concern. The embedding approach appears to be more sensitive to this distinction, scoring less for benevolence when the surrounding context does not support it.

Taken together, this example suggests a general trade-off. The dictionary is transparent, since every value assignment traces to a specific word, but it is vulnerable to false negatives (values expressed through theme rather than keywords, as with stimulation and security) and false positives (surface matches that ignore context, as with benevolence). This may be amplified in children's reviews, where

informal spelling and different vocabulary further reduce keyword matches. This example illustrates how values are often expressed implicitly, but can be found within these reviews. For our method specifically, the example also indicates that the embedding-based approach is better suited to capturing indirect, context-dependent value expression, though this comes at the cost of interpretability, making it harder to trace value profiles back to specific textual evidence.

“They are such **inspiring** people and i loved the book!!!:) Bethany Hamilton is the true definition of **faith** and **courage**. After being attacked by a shark at the age of 14 in her hometown while surfing off of Tunnels Beach she was attacked by a 15 foot tiger shark and she lost her arm. she never let that put her down. with an enormous **faith** she has returned to pro surfing. the movie is great too!!! reading this with my **family** every night at dinner!!! This is a great book. It is very suspenseful but to say the truth it has a lame ending. This is a FANTASTIC book for ANYONE!! Mostly girls This is better than I **thought**. Jayfeather, Squirreelflight, Dovewing and Foxleap go on a journey to the tribe of rushing water to fulfill a prophecy. Icecloud continues her quest in the Dark Forest as a spy. Another cat joins the ranks in the Dark Forest and one dies. This sounds like a great book. I like the **Beauty** and the Beast story. Making it present day makes it sound like an awesome book! i **enjoyed** this book recommend to anyone i **love** this book! :) this is my least favorite in the series still good this is an great book for anyone This is a story about a boy and his sister that go and spend the summer with their aunt at her old Vermont hotel, they research the history of the hotel and its relation to ghosts This is a realllllly good book. DO NOT judge the book by its cover! I did. But after my **mom** made me read it it is realllllly great! This is my favorite book of the Erin Hunter books! this is definitely the most intense and **fun**! you’ll have to read IT SO I WONT SPOIL IT FOR YOU! :) It is a scary, intense **action** book. This is a very good book with lots of intense **action**.”

RQ2: To what extent do book recommendations represent children’s individual values, and how do they compare in terms of reflecting, amplifying, or suppressing those values?

Turning to the extent to which recommendations reflect these values, the clearest finding is not which algorithm amplifies or suppresses which value in isolation, but that all four strategies converge on the same directional bias regardless of how they generate recommendations. This suggests that the observed value distortions are driven less by algorithm-specific design or user characteristics and more by characteristics shared across the recommendation pipeline, such as the underlying book catalog or the way values are encoded in book descriptions. We see this as security is consistently over-represented relative to what children expressed, by roughly 6-10% in relative terms across algorithms, and tradition is consistently under-represented. Given that all different strategies pick up on the same signal, the source of this effect is unlikely to be just algorithm design.

One possible explanation is that book descriptions simply contain more security-related language than tradition-related language, causing any algorithm trained on this catalog to reproduce that imbalance. However, this explanation does not fit well with our coverage results: value coverage is roughly uniform across recommendation lists, with all ten values appearing at similar frequency regardless of algorithm. is roughly uniform across recommendation lists, meaning all ten values appear with similar frequency regardless of algorithm. Thus, the amplification and suppression effect appears only when recommendations are compared with what children expressed. The source of the mis-alignment may instead lie less in the catalog’s composition than in how children’s own expressed values compare to this evenly distributed set of recommendations, or in more subtle aspects of how security- and tradition-related content is represented and encoded.

Studies of primary-grade children’s books report that themes such as family unity, love, and helpfulness, which map onto Schwartz’s security and benevolence dimensions, are among the most commonly represented personal and social values, whereas loyalty and frugality, which sit closer to the conservation values of tradition and conformity, are among the least common, alongside broader social themes

such as peace, empathy, and freedom, which correspond to universalism and self-direction [64]. This supports the idea that security- and benevolence-related themes are generally favored in children's literature, even if our coverage results suggest the effect we observe is not driven by volume alone. For children, this means that recommendations may repeatedly expose them to a particular subset of values regardless of what they themselves express or prioritize. A child whose reviews emphasize curiosity, independence, or tradition may still encounter recommendations that consistently promote themes of safety and protection, just because these are more prevalent in the underlying catalog. From the child's perspective, this shapes not only which books are recommended, but also which kinds of stories and value perspectives become most visible over time.

Coverage vs individual alignment Another finding is the divergence between catalog-level coverage and individual-level alignment. All four algorithms achieve normalized entropy scores above 0.99 for the distribution of values in recommendations, indicating that no value category is neglected or strongly propagated. Self-direction and stimulation in particular are well represented in this coverage measure. Yet both of these values are systematically suppressed once recommendations are compared against what individual children expressed. Viewed only through value coverage of recommended items, the recommendation system appears to represent all values adequately. In contrast, our individual-level alignment analysis reveals systematic patterns of value amplification and suppression that remain hidden when evaluation relies solely on aggregate metrics. This corresponds to a broader methodological concern raised in other RS research as well, namely that aggregate summaries can obscure precisely the group- or individual-level effects an evaluation is meant to surface, whether by engagement group in news RS [84] or, as argued for information retrieval systems more generally, by any relevant subgroup [28]. In practice, this means two children using the same system may receive recommendations that differ substantially in how well they resonate with the values they express, even though the system appears fair or representative when only the population level is considered.

Looking at algorithms individually adds a further layer of nuance. Most-Popular produces the single largest amplification, namely for security, and one of the largest suppression, tradition, yet it achieves the lowest mean per-child distortion of the four algorithms (0.2094, versus 0.2312-0.2497 for the others). This is not a contradiction, as distortion aggregates squared differences across all ten value categories. Most-Popular concentrates its mismatch on two extreme value-category pairs (security and tradition), while staying close to children's expressed priorities on the remaining eight values.

ItemKNN, by contrast, shows no single extreme mismatch of this kind, but deviates moderately from children's expressed values across a wider set of categories, e.g., achievement, universalism and hedonism, which adds up to a higher total distortion score despite no individual value being as distorted as security or tradition are for Most-Popular. The distortion metric seems to favor Most-Popular slightly, so it would give a lower distortion score to a recommendation list that is somewhat aligned to the child on nearly all values except two, compared to a list that is subtly off across the board.

Segment-level differences At the user-segment level, the picture becomes even more heterogeneous. Security and power remain the most amplified values in every segment, reinforcing that this is a structural rather than segment-specific effect, but the relative magnitude shifts: in the already conservation-leaning segment B, power is amplified more strongly than security, possibly because the segment expresses security more, leaving less room for further amplification. It is worth noting that a smaller, less expressive segment, such as B, has fewer members and a narrower value range to begin with; both factors can inflate or flatten apparent amplification relative to larger, more heterogeneous segments. Furthermore, tradition is no longer uniformly suppressed once segments are considered; this shows that a similar population mean does not imply similar behavior for a specific group of children.

Thus, some children encounter recommendations that align relatively closely with the values they express, while others systematically receive recommendations that are less representative of their interests and perspectives. As a result, children who are already less typical or less represented in the data may also be those whose values are least well reflected by the recommender. This finding closely relates to evidence from the news domain, where offline evaluation on a full user population was shown to obscure how differently distinct engagement-based clusters are served, motivating the argument that a single recommender strategy is unlikely to serve all clusters simultaneously [84]. Our results provide

an analogous demonstration in the domain of children’s book recommendation, for value-alignment.

While the amplification of security and power persists across all segments, other value distortions vary considerably between groups. Likewise, algorithm performance is not consistent across segments: methods that appear similarly effective when evaluated over the full population can differ substantially for individual groups of children. These findings show that population-level averages mask important differences in how recommendation algorithms affect distinct user segments, and that no single algorithm provides the best balance between accurate recommendation and value alignment for all children.

Accuracy vs alignment Finally, no measurable relationship was observed between recommendation accuracy and our value-alignment metrics. Spearman correlations between $nDCG@10$ and both amplification magnitude and distortion were negligible in magnitude, below 0.02, and not statistically significant for any of the four algorithms. The absence of correlation, alongside the fact that $nDCG@10$ itself is low and the underlying data sparse (top-10 hit rates ranging from 7.66% to 15.83%) limits the statistical power available to detect a modest relationship. This suggests that pursuing value alignment in this setting does not appear to require sacrificing standard ranking accuracy, reinforcing existing arguments that accuracy alone is an incomplete lens through which to evaluate recommender purpose [42, 76].

This absence of a population-level trade-off does not mean each algorithm treats every segment equally well. When we look at which algorithm each segment is best served by, only the largest and most homogeneous segment, the selective, openness-to-change-leaning majority, is well served on both accuracy and value alignment at once. For the remaining, smaller and more distinctive segments, the two measures pull apart, and in opposite directions: one segment is comparatively well aligned with the values it expresses but poorly served in terms of ranking accuracy, while another is served relatively well by accuracy but poorly by alignment. Rather than a single trade-off that applies uniformly across children, accuracy and value alignment appear to trade off differently depending on which group of children is considered, and only coincide for the segment that already resembles the implicit “average user” most closely.

Even where an algorithm does emerge as the best available option for a given segment, this does not mean it serves that segment well, only that it is the least poor fit among the alternatives tested. The smallest and least expressive segment, for instance, is best served by the same algorithm that also produces the strongest suppression of its most characteristic value at the population level. Its apparent advantage for this group may be an artifact of how neighborhood-based recommendation happens to connect similar, under-expressive, users to one another, rather than evidence that the algorithm is attentive to this segment’s values. Similarly, segments that look alike in how broadly they express values are not served alike by the same algorithms, suggesting that surface-level similarity in a segment’s value profile is not sufficient to predict how well any given recommender will serve it. These patterns suggest that “best-performing for a segment” is a comparative standard, not an absolute one, and that none of the four strategies considered here offers a dependable way of serving children segments well.

These findings reinforce the conclusion from RQ1 that children are not a homogeneous user group. Consistent with work arguing that children’s value priorities are shaped by their social, cultural, and political context, and that childhood itself is not a single, homogeneous category [2], we show children are individuals with different value preferences, rather than as instances of a generic child user. Age is one recognized factor of this variation: the literature on children’s value development suggests that openness-to-change values, including independence, tend to become more pronounced across middle childhood and into adolescence, relative to conservation values [21]. This is, however, a population-level trend rather than a rule that applies uniformly to every child. Because children process and respond to online content differently from one another and not only from adults, recommendations cannot simply be adapted for children, but must account for the within-group heterogeneity discussed here.

5.1. Implications for Children

Our findings suggest that children do not simply receive different book recommendations because they enjoy different stories. Rather, they may also experience systematically different representations of

personal values. Two children using the same recommender system may repeatedly encounter different emphases in the values embedded within recommended books, depending both on their own value profile and on the biases present in the algorithms or book catalogue. Consequently, RS contribute to shaping not only what children read, but also the kinds of perspectives and value-laden narratives they are exposed to daily.

Importantly, these effects are unlikely to be experienced equally by all children. Our clustering analysis shows that children differ not only in which values they express, but also in how strongly they express them. Consequently, the same structural biases in RS are likely to affect children differently. A child whose expressed values already align more closely with the values that RS systematically favor, such as security or power, is likely to encounter recommendations that continue to reflect those values. In contrast, children whose expressed values differ from this pattern, for example those leaning more toward tradition or belonging to smaller, less expressive groups such as Segment B, are comparatively less well represented by the recommendations they receive. In practice, this means that the children whose value profiles differ most from the implicit favored profile, the one built into the recommender, may also experience the greatest mismatch between the values they express and those embedded in their recommended content. If such uneven exposure accumulates over time, children may gradually end up in quite different digital reading environments, despite using the same system.

Although our study cannot determine whether recommendation exposure changes children's values, developmental literature suggests that middle childhood is a period during which value priorities remain relatively malleable [21, 24]. Systematic amplification or suppression of particular values may become more consequential than it would be for adults whose value structures are comparatively stable. This possibility reinforces the importance of understanding RS not only as tools for information access but also as environments in which children encounter recurring social and moral messages.

From this perspective, value alignment is not merely an optimization objective for RS, but a consideration for children's well-being in the digital space. A system that consistently over-represents some values while marginalizing others is implicitly shaping the environment in which children discover books, stories, and perspectives. Whether designers ultimately aim to reflect children's existing values, broaden their exposure, or balance multiple stakeholder perspectives, these choices should be made deliberately rather than emerge unintentionally from catalog composition or algorithmic bias.

5.2. Implications for RS Design

The most important practical implication is that children should *not* be treated as a single, undifferentiated unit for recommender design or evaluation. In practice, a recommender calibrated to, or evaluated for, "children" risks performing well for whichever segment happens to dominate the aggregate metric, while systematically under-serving smaller or less expressive groups, whose members remain largely invisible in population-level averages. For these children, poor value alignment is not an occasional mismatch but the consistent experience across the alternatives we tested. This occurs precisely because these are the children least likely to leave enough of a signal to be noticed in the first place. Thus, children are not as a single, homogeneous user group for RS.

Segment-aware recommendations. One concrete response is to condition recommendation and evaluation on which group of children a user belongs to, rather than optimizing a single algorithm against the population as a whole. For instance, routing a child to whichever algorithm best preserves their desired segment's values, or re-ranking a shared recommendation list to reduce segment-level distortion. This depends on reliably inferring segments in an online setting. Evaluation practice would need to follow, reporting alignment and accuracy per segment rather than only as a population-wide average, so that aggregate fairness cannot mask consistently poor treatment of a specific group. This is consistent with broader calls to move recommender evaluation and design beyond a single accuracy-driven objective toward purpose-specific design [42].

Together, RQ1 and RQ2 point to a finding with consequences beyond this dataset, namely that **recommendation algorithms are not neutral with respect to children's values**, even when no one designed them to favor any value at all. The security-amplified, tradition-suppressed pattern observed across every algorithm we considered is a form of implicit value curation, invisible to standard accuracy

evaluation. In practice, the bias is likely to appear regardless of which algorithm we use, but it suggests that these systems favor certain values over others, thus embedding value preferences without being designed to do so. This is precisely the kind of unacknowledged value embedding that motivates calls for value-sensitive design in RS research more broadly, in which values are considered explicitly in design rather than be left as side effects [32, 77, 61]. Our results provide concrete evidence that these concerns also apply to children’s book recommendation.

This raises an open normative question that we cannot settle simply on the basis of our empirical analysis: *should a recommender system reflect the values a child has already expressed as closely as possible, or should it deliberately expose the child to values they have not yet expressed, while those values are still forming?* The developmental literature reviewed for this thesis, discussed in Chapter 2, suggests children’s value priorities are not yet stable, with evidence of directional shifts, specifically an increase in openness-to-change and a decrease in conservation values, across middle childhood [21], a pattern broadly consistent with the general observation that value priorities show more change earlier in life and greater stability only in adulthood [66]. If repeated exposure to particular value-laden content genuinely contributes to value formation during this period, then a system that happens to track the average child’s profile reasonably well, while systematically over-exposing every child to security-themed content and under-exposing them to tradition-themed content, is not a neutral choice, even if no one intended to make it.

Engagement beyond our empirical evaluation is needed to answer this normative question and connects to a longer-running debate in children’s digital-rights, between protection-oriented and autonomy-oriented framings of what children need from online systems. Our findings show that current systems already make this choice implicitly rather than deliberately. Whichever stance a designer or organization ultimately adopts, the amplification, distortion, and coverage metrics used in our frameworks offer a concrete way to probe whether a system’s actual behavior matches that stance, in the same way fairness auditing is used to check whether a system’s actual behavior matches a fairness commitment [53]. However, auditing is not straightforward to operationalize in practice, as experience with fairness in production recommender systems documents conceptual difficulty, including disagreement over which definitions to apply and a persistent gap between statistical significance and meaningful harm [8].

Navigating this question would also benefit from directly involving children, caregivers, and educators as stakeholders, consistent with the more recent evolution of VSD away from prescriptive, top-down value orderings and toward participatory methods [11]. However, this is not easy in practice, as value tensions are shown to arise between different stakeholders. Related analysis of preschool story and tale books similarly finds that teachers, parents, and principals do not converge on a single set of value themes to emphasize when selecting or producing children’s reading material [15].

Recommender systems research has long recognized the risk of narrowing what users are exposed to, developing mechanisms such as diversity, novelty, and serendipity specifically to counteract overspecialization and the reinforcement of a user’s existing preferences. Within value-aware recommender research specifically, diversity, understood mainly as topic or viewpoint diversity, is by far the most frequently studied value, while other value dimensions remain comparatively neglected, particularly within computer science [7]. The value-alignment concern raised by us sits within this same broader tradition of guarding against narrowing exposure, but targets a different, unexplored lens: which personal values are amplified or suppressed relative to what they have already expressed.

5.3. Limitations and Future Work

While our approach yields meaningful insights into how RS reflect children’s values, it also comes with several limitations, both in the data used and in the framework itself, which we discuss below alongside directions for future work.

Value framework. We operationalize “values” exclusively through Schwartz’s ten basic values, and our findings on children’s expressed values, on user segments, and on recommender alignment are dependent on this specific taxonomy. This may not fully capture childhood-specific values ([18]) and the most recently validated framework of Schwartz’s values, the refined theory of basic human values ([69]), introduced finer distinctions to discriminate between 19 value types. We nonetheless adopt Schwartz’s basic theory because it is the most cross-culturally validated value taxonomy, which has

been incorporated in studies with children as well. Applying alternative value taxonomies to the same pipeline, or empirically comparing how well different frameworks fit children’s book reviews, is a possible direction for future work.

Reviews as a proxy for value expression. Our pipeline relies on children’s written reviews, which restricts our sample to children who read *and* chose to write about what they read. Such users may be those who are more engaged and possibly older or more literate subset of the broader population of child readers. This filtering step was necessary, since reviews are the only available window into expressed values in this dataset and text written freely by people seems to be a more reliable indicator of authentic values than self-report questionnaires [13], a reliability gap that is documented to be even larger for children specifically [35], supporting reviews as a suitable medium for this population. Still, the size of this filtering effect is worth noting explicitly. A natural extension is exploring whether other signals, such as genre or rating patterns, could extend value inference to children who read but do not write reviews. For example, a preference for certain genres, such as fantasy over comics, may give additional insights into underlying values, but was not the purpose of our research, where we focus on children’s value expression rather than inference of these values.

Child-user identification via a genre proxy. In the absence of demographic data, we identified likely child users based on disproportionate engagement with children’s-genre books, following established practice for this kind of inference [80, 81]. In the absence of demographics, this is a widely used approach, but it introduces measurement error in both directions: some identified users are plausibly adults engaging with children’s literature, such as parents or teachers, while some child readers who do not concentrate heavily on the children’s genre may be excluded. While a strong concentration in children’s genre is itself informative, suggesting these values may be transmitted in children’s lives, our results should be seen in light of this. Family remains one of the primary channels through which value orientation is shaped in children [33], lending some plausibility to treating adults with heavy engagement in children’s literature, such as parents, as informative for this analysis even when they are not children themselves.

Interaction filtering. Our pre-processing pipeline (retaining only reviewed, positively-rated interactions from users with a minimum interaction count) substantially reduces the size of the dataset. Each step is individually justified: reviews are required for text-based value signal, positive-only interactions match how modern recommenders are typically evaluated, and a minimum interaction count is needed for reliable value profiles, particularly important for our offline study. These steps also align our setup with common practice in the recommender systems literature, making our results comparable to prior work. Still, the cumulative effect of applying all three is likely to bias the analyzed sample toward more engaged, more expressive users to begin with. Data availability for children is a recognized structural constraint across children-centered RS research, driven in part by the data-protection requirements owed to minors, which inherently limit both dataset scale and the granularity of available signal [35]. Observational approaches to studying value expression could help uncover patterns among users excluded by this filtering.

Value detection. The value-detection approach itself, comparing sentence embeddings of reviews against sentence embeddings of value definitions, was selected because it best reproduced Schwartz’s theoretical value structure in preliminary comparisons and because organic text has been argued to outperform self-report instruments for value measurement generally [13]. Nonetheless, we compress varied, context-dependent language into a single similarity score per value per user, a trade-off between scale and interpretability noted in the value-detection literature when comparing automated approaches to manual coding [60]. Moreover, the validity evidence available for embedding-based value detection in the literature is drawn from adult-generated text, and it is not established that value-laden language operates identically in children’s book reviews, which differ from adult text in vocabulary and syntax. Yet, for our analysis, it provides an appropriate proxy, offering a scalable and theoretically grounded way to surface value signals across a large corpus of reviews. The example presented in 5 provides additional context to how our value detection method works and its benefits over traditional dictionary-based approaches.

Moreover, it is worth noting that most sub-clusters that emerge and recommendations are value-coherent with respect to the Schwartz value model. Elevated dimensions within a sub-cluster tend to sit adjacent to one another on the Schwartz circumplex (e.g. security and tradition co-occurring, or

self-direction and stimulation) rather than combining values Schwartz’s theory treats as being in tension. Because Schwartz’s circular structure has been substantially replicated across cultures using self-report instruments, this adjacency serves as a sanity check that the value-detection pipeline is picking up on a coherent signal in children’s language rather than noise. It also extends existing evidence that organic, self-generated text can carry theoretically consistent value structure [13, 16]. We also observe that generally, algorithms do not amplify or suppress conflicting values, but rather favor one dimension over others.

With this validity check in mind, the compression of values remains a limitation in a more conceptual sense. Recent ethnographic work on algorithmic systems describes “value narrowing”: the process by which a broad, culturally embedded value is converted into something technical and measurable, while discussion of it continues to invoke the value’s wider meaning [57]. Our pipeline is susceptible to a version of this: tradition, for Schwartz, spans respect for and commitment to one’s cultural or religious customs, yet we measure it as similarity to a single, static definition sentence, which cannot distinguish, for example, family ritual from religious practice or general nostalgia. This narrower, technical proxy is what allows the value to be measured consistently at scale, but it also means our findings speak to this specific operationalization rather than to the full breadth of what tradition, or any Schwartz value, can mean for a child. Enriching value definitions with more context-specific or culturally grounded formulations, potentially varying by sub-population, is a promising direction for making value detection more sensitive to this diversity in future work.

This tension between measurement granularity and scale also intersects with the temporal dimension of our analysis. Our value profiles represent a single snapshot of each child’s expressed values, whereas developmental research indicates children’s value priorities shift over time before stabilizing [21]. Longitudinal observations of both expressed values and recommendation exposure over time is a natural direction for future work, and one that could be combined with the more context-sensitive value definitions discussed above.

A related constraint concerns how we evaluate recommendations themselves. Our alignment and accuracy metrics are computed against held-out historical interactions rather than against recommendations children were actually shown and reacted to. This is standard practice in RS research and appropriate for an exploratory analysis, but it means our findings describe what a system *would* recommend under these algorithms, not evidence of how children actually perceive or respond to the amplification patterns we observe. User studies assessing whether children notice or are affected by these patterns would be a valuable complement to the offline analysis presented here.

Future work could also apply this framework to other recommendation domains that children engage with, such as social media, video platforms, or music streaming. Nothing in the framework itself, deriving value profiles from user-generated text and item metadata, then comparing them per individual and per segment, is specific to books: for social media, a user’s comments could form the basis of their value profile, with captions serving the same role for recommended content; for music, lyrics could be used to profile recommended items. Applying the same approach across other domains could reveal whether the amplification and suppression patterns observed here for books are specific to this content type and catalog, or whether they reflect a more general property of RS, in which case similar value distortions might be expected across the broader range of platforms children use daily. This extension is particularly worth pursuing given that children often spend more time on video and social platforms than on reading apps, making them equally important environments to understand in terms of repeated value exposure. Since different value orientations have been linked to different developmental outcomes, for example, stronger self-enhancement values are associated with greater aggression in some adolescents, while self-transcendence values relate to more prosocial behavior [24], extending this analysis to the domains children engage with most could help investigate whether systematic amplification or suppression of particular values connects to these broader outcomes.

Taken together, these limitations do not undermine the core findings of our research, but help define our scope: an empirically grounded, reproducible framework for surfacing value (mis)alignment in recommender systems for children, built on the most complete data reasonably available without direct access to child participants. We see this as a foundation other researchers can extend to other domains, other taxonomies, other populations, and eventually, to the live systems children actually encounter, rather than a finished account of how recommenders and children’s values interact.

CONCLUSION

This manuscript set out to investigate whether RS reflect the values children express in one of the most formative media content, namely books. Motivated by the recognition that RS are not a neutral technology with respect to the values they embed and prioritize, we asked what values children prioritize in their book reviews, and to what extent book recommendations reflect, amplify, or suppress these values.

Our findings show that children’s value expression, captured through an embedding-based analysis of book reviews grounded in Schwartz’s theory of basic values, is diverse. At the population level, values appear moderately and evenly expressed, but clustering reveals four distinct user segments differing primarily in expressiveness and secondarily in value orientation, and further sub-cluster analysis exposes heterogeneity concealed within these broader groups. Turning to recommendation alignment, we found that four collaborative-filtering algorithms, despite differing in design, converge on the same directional bias: security is consistently amplified and tradition consistently suppressed relative to what children expressed, a pattern that persists across user segments and appears to originate in catalog-level characteristics rather than algorithmic choice. For systems used by children, whose values are still developing, such recurring patterns of exposure warrant deliberate consideration rather than being left as unintended consequences of technical optimization. Crucially, this misalignment is invisible under standard aggregate coverage metrics, which suggest values are well represented.

The central contribution of this thesis is twofold. First, it offers an empirically grounded, reproducible framework for detecting values in children’s textual expressions without relying on direct self-report and for quantifying value alignment between users and recommendations. Second, it provides an empirical demonstration that RS are not value-neutral with respect to children specifically, and that this non-neutrality operates unevenly across the child population.

These findings have practical implications for RS design and evaluation. Treating children as a single, undifferentiated user group risks hiding meaningful disparities in whose values are presented. It shows that value misalignment can exist as a systematic property of a recommendation pipeline and that such misalignment can remain unnoticed under conventional accuracy metrics. For RS research, this underscores the need for segment- and individual-level evaluation practices, and it extends the value-sensitive design literature by demonstrating a concrete methodology for assessing value representation empirically rather than only theoretically. For practitioners, it raises a genuine design question, largely unaddressed by current systems, of whether recommenders should mirror children’s existing values or deliberately diversify their exposure during a developmental period when values are still forming.

Ultimately, our work highlights a need to look beyond the individual child as a data point and toward the individuals within: the values they express, and the values a system chooses to reflect back to them. We open the floor for a broader discussion that the field has yet to take up: should recommenders mirror the values children already hold, or actively diversify what they encounter while those values are still forming?

REFERENCES

- [1] Berke Akkaya. “Current trends in recommender systems: A survey of approaches and future directions”. In: *Computer Science* 10.1 (2025), pp. 53–91.
- [2] Meryl Alper, Vikki S Katz, and Lynn Schofield Clark. “Researching children, intersectionality, and diversity in the digital age”. In: *Children, Adolescents, and Media*. Routledge, 2018, pp. 129–136.
- [3] Vito Walter Anelli et al. “Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation”. In: *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. Ed. by Fernando Diaz et al. ACM, 2021, pp. 2405–2414. DOI: 10 . 1145 / 3404835 . 3463245. URL: <https://doi.org/10.1145/3404835.3463245>.
- [4] Alissa N Antle. “Child-based personas: need, ability and experience”. In: *Cognition, Technology & Work* 10.2 (2008), pp. 155–166. URL: <https://doi.org/10.1007/s10111-007-0071-2>.
- [5] Oscar Araque, Lorenzo Gatti, and Kyriaki Kalimeri. “MoralStrength: Exploiting a moral lexicon and embedding similarity for moral foundations prediction”. In: *Knowledge-based systems* 191 (2020), p. 105184.
- [6] Oscar Araque, Ganggao Zhu, and Carlos A Iglesias. “A semantic similarity-based perspective of affect lexicons for sentiment analysis”. In: *Knowledge-Based Systems* 165 (2019), pp. 346–359.
- [7] Christine Bauer et al. “Where are the values? a systematic literature review on news recommender systems”. In: *ACM Transactions on Recommender Systems* 2.3 (2024), pp. 1–40.
- [8] Lex Beattie, Dan Taber, and Henriette Cramer. “Challenges in translating research to practice for evaluating fairness and bias in recommendation systems”. In: *Proceedings of the 16th ACM Conference on Recommender Systems*. 2022, pp. 528–530.
- [9] Tilde Bekker et al. *Editorial special issue on assumptions about the concept of childhood and the roles of children in design*. 2019.
- [10] Bruno Bettelheim. *The uses of enchantment: The meaning and importance of fairy tales*. Thames and Hudson London, 1976.
- [11] Alan Borning and Michael Muller. “Next steps for value sensitive design”. In: *Proceedings of the SIGCHI conference on human factors in computing systems*. 2012, pp. 1125–1134.
- [12] Gloria Swindler Boutte, Ronnie Hopkins, and Tyrone Waklatsi. “Perspectives, voices, and world-views in frequently read children’s books”. In: *Early education and development* 19.6 (2008), pp. 941–962. URL: <https://doi.org/10.1080/10409280802206643>.
- [13] Ryan Boyd et al. “Values in words: Using language to evaluate and understand personal values”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 9. 1. 2015, pp. 31–40.
- [14] Cletus R Bulach and Judy D Butler. “The occurrence of behaviors associated with sixteen character values”. In: *The Journal of Humanistic Counseling, Education and Development* 41.2 (2002), pp. 200–214.
- [15] Şerife Cengiz and Erol Duran. “Analysis of values on preschool period children story and tale books”. In: *Education and Science* 42.191 (2017). URL: <https://doi.org/10.15390/EB.2017.6945>.
- [16] Jilin Chen et al. “Understanding individuals’ personal values from social media word use”. In: *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 2014, pp. 405–414.

- [17] Fabian Christoffel et al. "Blockbusters and wallflowers: Accurate, diverse, and scalable recommendations with random walks". In: *Proceedings of the 9th ACM Conference on Recommender Systems*. 2015, pp. 163–170.
- [18] Jan Cieciuch, Eldad Davidov, and René Algesheimer. "The stability and change of value structure and priorities in childhood: A longitudinal study". In: *Social Development* 25.3 (2016), pp. 503–527.
- [19] Patricia R Collins et al. "Examining the consistency and coherence of values in young children using a new Animated Values Instrument". In: *Personality and Individual Differences* 104 (2017), pp. 279–285.
- [20] Ana Claudia Da Costa, Francisco Rebelo, and Júlia Teles. "Child-persona: What i think to what they are". In: *Advances in Ergonomics in Design: Proceedings of the AHFE 2016 International Conference on Ergonomics in Design, July 27-31, 2016, Walt Disney World®, Florida, USA*. Springer. 2016, pp. 43–51.
- [21] Ella Daniel, Kinneret Misgav, and Anat Chomsky. "Middle childhood development in personal values". In: *Journal of Personality* 92.5 (2024), pp. 1283–1298. URL: <https://doi.org/10.1111/jopy.12888>.
- [22] Anna K Döring. "Assessing children's values: An exploratory study". In: *Journal of Psychoeducational Assessment* 28.6 (2010), pp. 564–577.
- [23] Anna K Döring, Ella Daniel, and Ariel Knafo-Noam. "Introduction to the special section value development from middle childhood to early adulthood—new insights from longitudinal and genetically informed research". In: *Social Development* 25.3 (2016), pp. 471–481.
- [24] Anna K Döring, Ella Daniel, and Ariel Knafo-Noam. "Introduction to the special section value development from middle childhood to early adulthood—new insights from longitudinal and genetically informed research". In: *Social Development* 25.3 (2016), pp. 471–481. URL: <https://doi.org/10.1111/sode.12177>.
- [25] Anna K Döring et al. "Assessing values at an early age: The picture-based value survey for children (PBVS-C)". In: *Journal of personality assessment* 92.5 (2010), pp. 439–448.
- [26] Anna K Döring et al. "Cross-cultural evidence of value structures and priorities in childhood". In: *British Journal of Psychology* 106.4 (2015), pp. 675–699.
- [27] Michael D Ekstrand et al. "Exploring author gender in book rating and recommendation". In: *Proceedings of the 12th ACM conference on recommender systems*. 2018, pp. 242–250.
- [28] Michael D Ekstrand et al. "Not just algorithms: Strategically addressing consumer impacts in information retrieval". In: *European Conference on Information Retrieval*. Springer. 2024, pp. 314–335.
- [29] Salma Elsayed-Ali et al. "Designing for children's values: conceptualizing value-sensitive technologies with children". In: *Proceedings of the 2020 ACM Interaction Design and Children Conference: Extended Abstracts*. 2020, pp. 296–301.
- [30] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *nature* 542.7639 (2017), pp. 115–118. URL: <https://doi.org/10.1038/nature21056>.
- [31] Katherine Hunt Federle. "Children's Rights and the Need for Protection". In: *Fam. LQ* 34 (2000), p. 421.
- [32] Batya Friedman et al. "Value sensitive design and information systems". In: *Early engagement and new technologies: Opening up the laboratory*. Springer, 2013, pp. 55–95. URL: https://doi.org/10.1007/978-94-007-7844-3_4.
- [33] Eva Frýdková. "School and parents' cooperation on formation of value orientation of a child". In: *Procedia-Social and Behavioral Sciences* 106 (2013), pp. 2307–2315.
- [34] Iason Gabriel. "Artificial intelligence, values, and alignment". In: *Minds and machines* 30.3 (2020), pp. 411–437. URL: <https://doi.org/10.1007/s11023-020-09539-2>.
- [35] Emilia Gómez Gutiérrez, Vicky Charisi, and Stephane Chaudron. "Evaluating recommender systems with and for children: towards a multi-perspective framework". In: *CEUR Workshop Proceedings*. 2021; 2955. CEUR Workshop Proceedings. 2021.
- [36] Neil L Griffiths et al. "The values of only-children: Power and benevolence in the spotlight". In: *Journal of Research in Personality* 92 (2021), p. 104096.

- [37] Lucien Heitz. "Classification of Normative Recommender Systems." In: *NORMALize@ RecSys*. 2023.
- [38] Natali Helberger. "On the democratic role of news recommenders". In: *Algorithms, automation, and news*. Routledge, 2021, pp. 14–33.
- [39] Sybren Heyndels. "Technology and neutrality". In: *Philosophy & Technology* 36.4 (2023), p. 75. URL: <https://doi.org/10.1007/s13347-023-00672-1>.
- [40] Steven Hitlin and Jane Allyn Piliavin. "Values: Reviving a dormant concept". In: *Annu. Rev. Sociol.* 30.1 (2004), pp. 359–393.
- [41] Yinghui Huang et al. "Can psychographics mitigate over-specialization in recommender-driven consumer markets? Evidence from recommender systems based simulation experiment". In: *Management System Engineering* 4.1 (2025), p. 20.
- [42] Dietmar Jannach and Gediminas Adomavicius. "Recommendations with a purpose". In: *Proceedings of the 10th ACM conference on recommender systems*. 2016, pp. 7–10.
- [43] Hatice Kapusuz. "Ideology in Children's Books: Values Education Against Child Autonomy". MA thesis. Middle East Technical University (Turkey), 2024.
- [44] Mohammed Khwaja et al. "Aligning daily activities with personality: towards a recommender system for improving wellbeing". In: *Proceedings of the 13th acm conference on recommender systems*. 2019, pp. 368–372. URL: <https://doi.org/10.1145/3298689.3347020>.
- [45] Bart P Knijnenburg, Saadhika Sivakumar, and Daricia Wilkinson. "Recommender systems for self-actualization". In: *Proceedings of the 10th acm conference on recommender systems*. 2016, pp. 11–14. URL: <https://doi.org/10.1145/2959100.2959189>.
- [46] Kun Lin et al. "Calibration in collaborative filtering recommender systems: a user-centered analysis". In: *Proceedings of the 31st ACM Conference on Hypertext and Social Media*. 2020, pp. 197–206.
- [47] Enrico Liscio et al. "Axes: Identifying and Evaluating Context-Specific Values." In: *AAMAS*. 2021, pp. 799–808.
- [48] Sonia Livingstone et al. "Risks and safety on the internet: the perspective of European children: full findings and policy implications from the EU Kids Online survey of 9-16 year olds and their parents in 25 countries". In: (2011). URL: <https://eprints.lse.ac.uk/33731/>.
- [49] CL Lloreda. "Speech recognition tech is yet another example of bias". In: *Scientific American* 5 (2020). URL: <https://www.scientificamerican.com/article/speech-recognition-tech-is-yet-another-example-of-bias/>.
- [50] Bernard Mallia. "Designing and Implementing Socially Beneficial Recommender Systems: An Interdisciplinary Approach." In: *NORMALize@ RecSys*. 2023.
- [51] Leland McInnes, John Healy, Steve Astels, et al. "hdbscan: Hierarchical density based clustering." In: *J. Open Source Softw.* 2.11 (2017), p. 205.
- [52] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).
- [53] Silvia Milano, Mariarosaria Taddeo, and Luciano Floridi. "Recommender systems and their ethical challenges". In: *Ai & Society* 35.4 (2020), pp. 957–967.
- [54] Boaz Miller. "Is technology value-neutral?" In: *Science, Technology, & Human Values* 46.1 (2021), pp. 53–80. URL: <https://doi.org/10.1177/0162243919900965>.
- [55] Boaz Miller and Isaac Record. "Responsible epistemic technologies: A social-epistemological analysis of autocompleted web search". In: *New Media & Society* 19.12 (2017), pp. 1945–1963. URL: <https://doi.org/10.1177/1461444816644805>.
- [56] Joanna Misztal-Radecka and Bipin Indurkha. "Persona prototypes for improving the qualitative evaluation of recommendation systems". In: *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 2020, pp. 206–212. URL: <https://doi.org/10.1145/3386392.3399297>.
- [57] David Moats et al. "The dark side of public values in algorithmic systems". In: *Dialogues on Digital Society* (2026), p. 29768640261457507.

- [58] David R Morrow. "When technologies makes good people do bad things: Another argument against the value-neutrality of technologies". In: *Science and Engineering Ethics* 20.2 (2014), pp. 329–343. URL: <https://doi.org/10.1007/s11948-013-9464-1>.
- [59] Julian Nida-Rümelin and Klaus Staudacher. "Philosophical foundations of digital humanism". In: *Introduction to digital humanism: A textbook*. Springer, 2023, pp. 17–30.
- [60] Vladimir Ponizovskiy et al. "Development and validation of the personal values dictionary: A theory-driven tool for investigating references to basic human values in text". In: *European Journal of Personality* 34.5 (2020), pp. 885–902.
- [61] Erich Prem. "Digital humanism and norms in recommender systems". In: *17th ACM Conference on Recommender Systems: First workshop on normative design and evaluation of recommender systems*. 2023.
- [62] Paul Resnick et al. "Grouplens: An open architecture for collaborative filtering of netnews". In: *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. 1994, pp. 175–186.
- [63] Francesco Ricci, Lior Rokach, and Bracha Shapira. "Recommender systems: Techniques, applications, and challenges". In: *Recommender systems handbook* (2021), pp. 1–35.
- [64] Ayfer Şahin. "Personal and social values in primary grade children's books". In: *Participatory Educational Research* 6.1 (2019), pp. 1–9. URL: <https://doi.org/10.17275/per.19.1.6.1>.
- [65] Badrul Sarwar et al. "Item-based collaborative filtering recommendation algorithms". In: *Proceedings of the 10th international conference on World Wide Web*. 2001, pp. 285–295.
- [66] Shalom H Schwartz. "An overview of the Schwartz theory of basic values". In: *Online readings in Psychology and Culture* 2.1 (2012), p. 11. URL: <https://doi.org/10.9707/2307-0919.1116>.
- [67] Shalom H Schwartz. "Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries". In: *Advances in experimental social psychology*. Vol. 25. Elsevier, 1992, pp. 1–65.
- [68] Shalom H Schwartz et al. "Extending the cross-cultural validity of the theory of basic human values with a different method of measurement". In: *Journal of cross-cultural psychology* 32.5 (2001), pp. 519–542.
- [69] Shalom H Schwartz et al. "Refining the theory of basic individual values." In: *Journal of personality and social psychology* 103.4 (2012), p. 663.
- [70] Reut Shachnai and Ella Daniel. "Children's concepts of personal values". In: *Infant and Child Development* 29.6 (2020), e2205.
- [71] Jin Shang, Mingxuan Sun, and Kevyn Collins-Thompson. "Demographic inference via knowledge transfer in cross-domain recommender systems". In: *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE. 2018, pp. 1218–1223.
- [72] Yiting Shen, Steven R Wilson, and Rada Mihalcea. "Measuring personal values in cross-cultural user-generated content". In: *International Conference on Social Informatics*. Springer. 2019, pp. 143–156.
- [73] Oscar Smallenbroek et al. "Constructing Schwartz values framework using the Rokeach values survey: Human value measurement in the longitudinal internet survey for social sciences". In: *PloS one* 20.8 (2025), e0329179.
- [74] Jonathan Stray. "Aligning AI optimization to community well-being". In: *International Journal of Community Well-Being* 3.4 (2020), pp. 443–463. URL: <https://doi.org/10.1007/s42413-020-00086-3>.
- [75] Jonathan Stray. "Aligning ai to human values means picking the right metrics". In: *Partnership on AI* 15 (2020). URL: <https://medium.com/partnership-on-ai/aligning-ai-to-human-values-means-picking-the-right-metrics-855859e6f047>.
- [76] Jonathan Stray et al. "Building human values into recommender systems: An interdisciplinary synthesis". In: *ACM Transactions on Recommender Systems* 2.3 (2024), pp. 1–57. URL: <https://doi.org/10.1145/3632297>.
- [77] Jonathan Stray et al. "What are you optimizing for? aligning recommender systems with human values". In: *arXiv preprint arXiv:2107.10939* (2021).

- [78] Mingxuan Sun, Changbin Li, and Hongyuan Zha. "Inferring private demographics of new users in recommender systems". In: *Proceedings of the 20th ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems*. 2017, pp. 237–244.
- [79] Robin Ungruh. "Are Recommender Systems Serving Children? Toward Child-Aware Design and Evaluation". In: *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. 2025, pp. 1451–1457.
- [80] Robin Ungruh, Murtadha Al Nahadi, and Maria Soledad Pera. "Mirror, mirror: exploring stereotype presence among top-n recommendations that may reach children". In: *ACM Transactions on Recommender Systems* (2025).
- [81] Robin Ungruh and Maria Soledad Pera. "This Is The World (Recommender Systems Made It For You) Capturing Risk Pathways for Children through a Pluralistic Multi-Risk Assessment of Recommenders". In: *Proceedings of the 34th ACM Conference on User Modeling, Adaptation and Personalization*. 2026, pp. 233–243.
- [82] Robin Ungruh et al. "Impacts of Mainstream-Driven Algorithms on Recommendations for Children Across Domains: A Reproducibility Study". In: *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. 2025, pp. 783–791.
- [83] Robin Ungruh et al. "When Preference Is Not Enough: Why Recommender Systems Require Human-Aware Evaluation for Children". In: *Proceedings of the 34th ACM Conference on User Modeling, Adaptation and Personalization*. 2026, pp. 563–566.
- [84] Hanne Vandenbroucke et al. "Not One News Recommender To Fit Them All: How Different Recommender Strategies Serve Various User Segments". In: *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. 2025, pp. 643–648.
- [85] Kurt Wagner. *Inside Twitter's ambitious plan to change the way we tweet*. 2019. URL: <https://www.vox.com/2019/3/8/18245536/exclusive-twitter-healthy-conversations-dunking-research-product-incentives>.
- [86] Mengting Wan and Julian McAuley. "Item recommendation on monotonic behavior chains". In: *Proceedings of the 12th ACM conference on recommender systems*. 2018, pp. 86–94.
- [87] Mengting Wan et al. "Fine-grained spoiler detection from large-scale review corpora". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 2605–2610.
- [88] Ge Wang et al. "Challenges and opportunities in translating ethical AI principles into practice for children". In: *Nature Machine Intelligence* 6.3 (2024), pp. 265–270.
- [89] Ge Wang et al. "Informing age-appropriate ai: Examining principles and practices of ai for children". In: *Proceedings of the 2022 CHI conference on human factors in computing systems*. 2022, pp. 1–29.
- [90] Udi Weinsberg et al. "BlurMe: Inferring and obfuscating user gender based on ratings". In: *Proceedings of the sixth ACM conference on Recommender systems*. 2012, pp. 195–202.
- [91] Hannes Werthner et al. *Perspectives on digital humanism*. Springer Nature, 2022.
- [92] Haoran Ye et al. "Measuring human and ai values based on generative psychometrics with large language models". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 25. 2025, pp. 26400–26408.
- [93] Bieke Zaman and Vero Vanden Abeele. "Laddering with young children in User eXperience evaluations: theoretical groundings and a practical case". In: *Proceedings of the 9th International Conference on Interaction Design and Children*. 2010, pp. 156–165.

ADDITIONAL FIGURES FOR RESULTS OF VALUE PROFILES

Value Profile Distributions

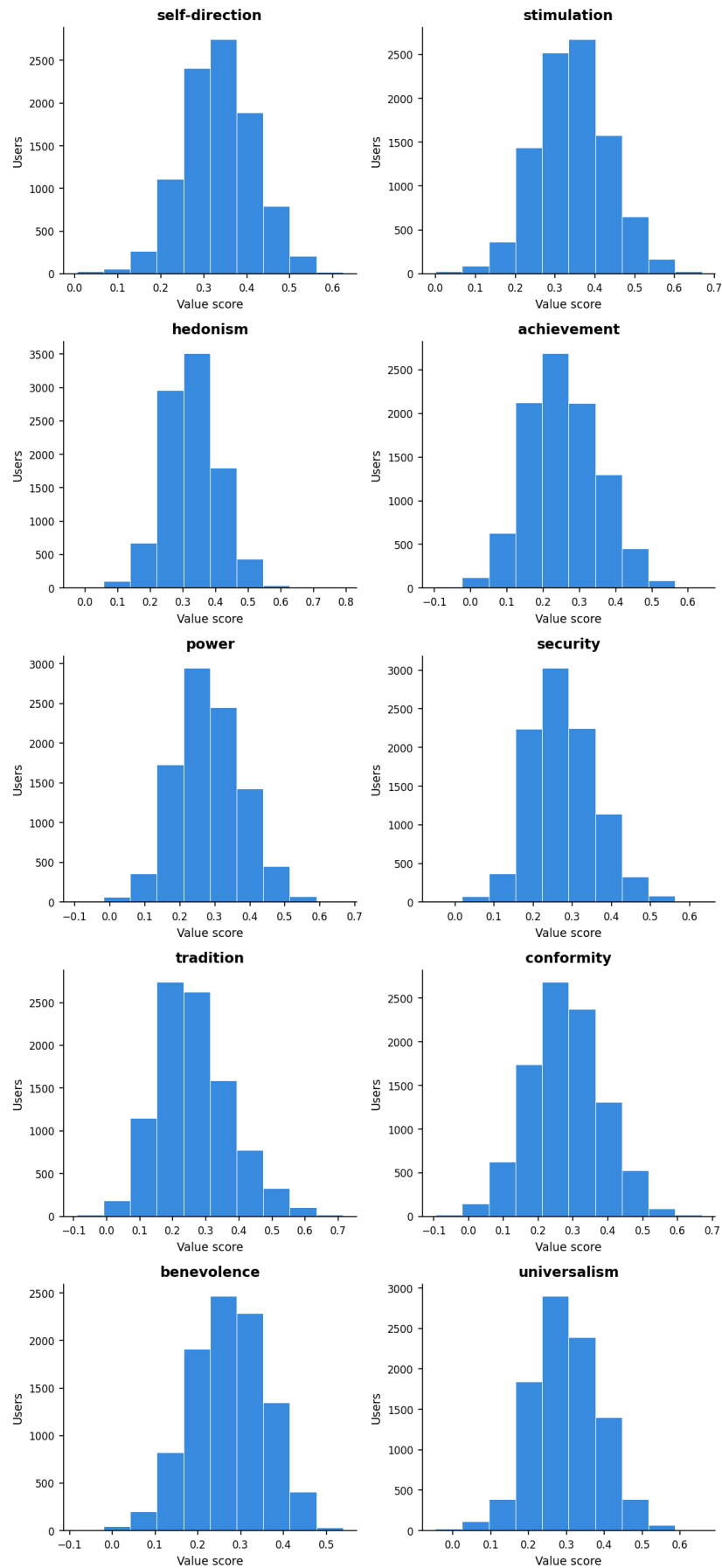


Figure A.1: Distribution of value scores over the 10 values.

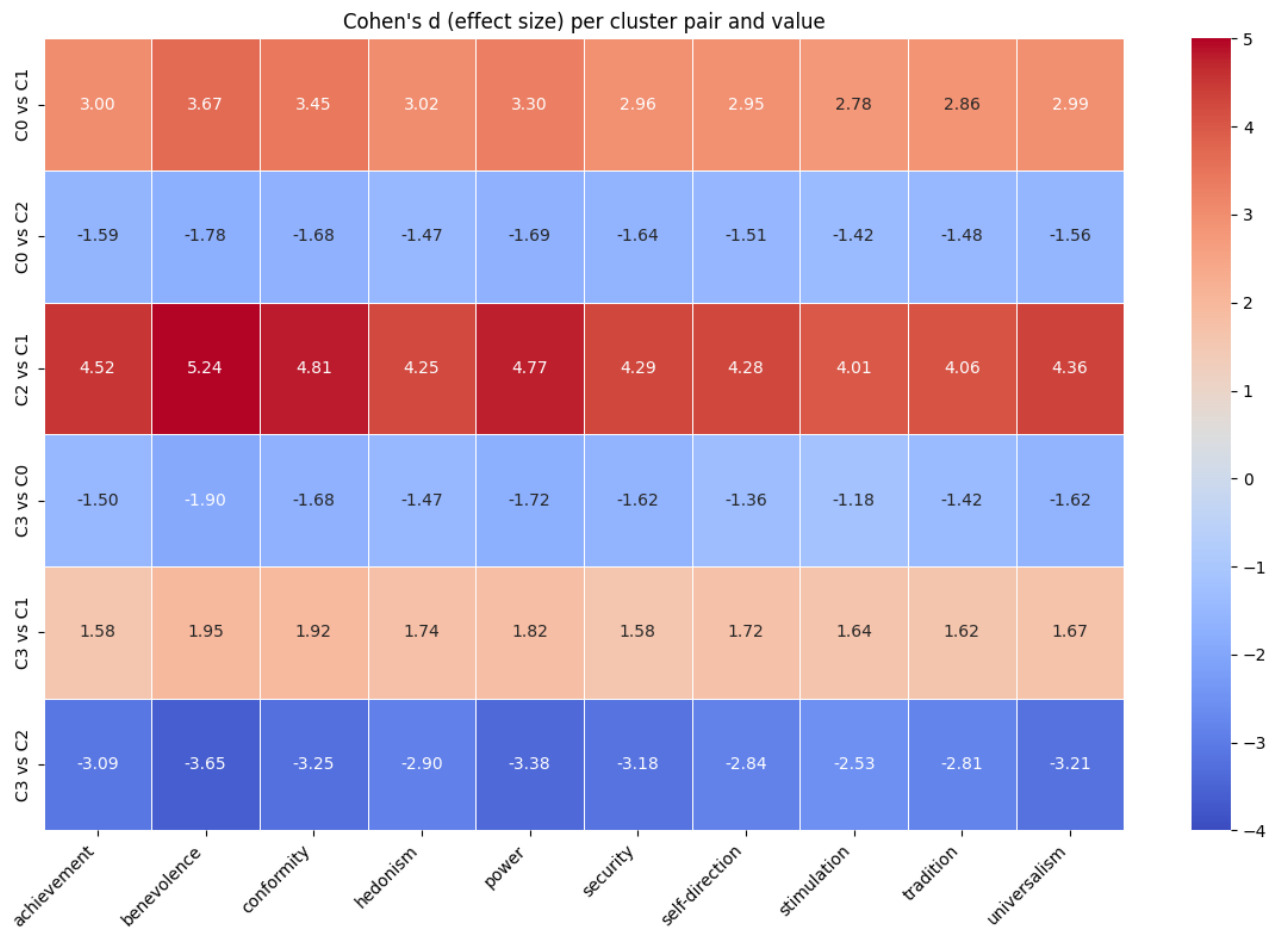


Figure A.2: Cohen's d value (effect size) per cluster pair and value dimension.