

Proposal to encode the Khojki letter QA in Unicode

Anshuman Pandey


pandey@umich.edu


pandey.github.io/unicode


May 21, 2021

1 Introduction




This is a proposal to encode a new character in the Khojki block of the Unicode standard:





GLYPH	CODE	CHARACTER NAME
	1123F	KHOJKI LETTER QA

This letter was described in “Final Proposal to Encode the Khojki Script in ISO/IEC 10646” (L2/11-021), but was not include in the repertoire due to insufficient information at that time (p. 13). Now, a decade later, I’ve had the occasion to analyze Khojki manuscripts and printed books that attest to the identity of  as a semantically distinct character. The evidence supports encoding the letter in Unicode, which will further enable the complete and accurate representation of Khojki documents in digital plain text.

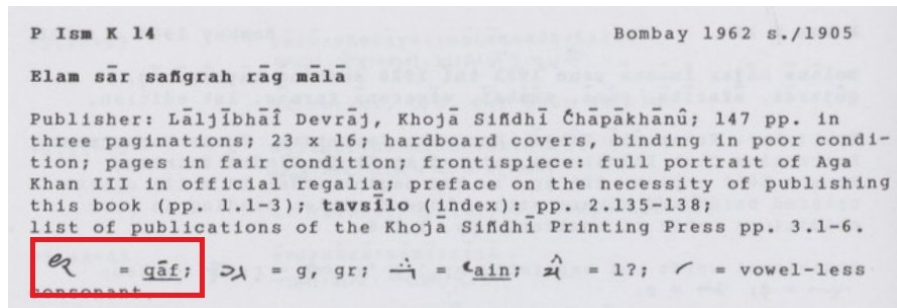
The proposed character name for  is KHOJKI LETTER QA, which references the primary sound value expressed by the letter. I have designed the representative glyph to conform with the letterforms used in the Khojki code chart, which are based on the ‘Khojki Jiwa’ font designed by Pyarali Jiwa.

2 Description

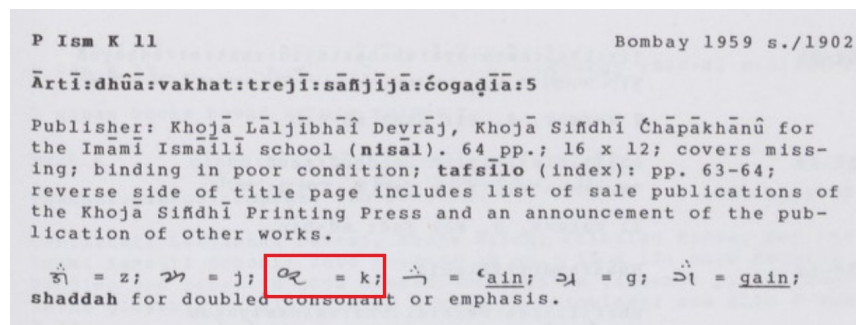
In Indic languages the voiceless uvular stop /q/ found in Arabic loanwords — represented by ق U+0642 ARABIC LETTER QAF — is pronounced using the voiceless velar stop /k/. Accordingly, when Arabic words are transliterated into Indic scripts, the /q/ is represented using the letter KA, which expresses /k/. In northern Indic scripts, the combining sign NUKTA may be written with KA to explicitly signal that the letter conveys a non-Indic sound akin to /k/, which readers generally interpret as the Arabic /q/. Khojki follows the same convention: the /q/ is commonly written as /k/ using  U+11208 KHOJKI LETTER KA or it may be specifically referenced by placing  U+11236 KHOJKI SIGN NUKTA above KA, ie. .

There is another, unique method for representing Arabic /q/ in Khojki. In several handwritten and printed documents, the letter  is used concurrently with  to differentiate /q/ and /k/. The origins of  are unknown. Its shape suggests a relationship to  ka, of which it may be a regional alternate or a historical

form. Indeed, **𐤀** and **𐤁** have the same archetype, distinguished by their initial strokes, but possessing the same central and terminal strokes. The **𐤁** is not described in Khojki script primers or shown in conventional charts of the script. But, it is documented in the scholarly literature, eg. in *The Harvard Collection of Ismaili Literature in Indic Languages* (p. 294), compiled by Ali S. Asani (1992):











Both **𐤁** *qa* and **𐤀** *ka* occur concurrently in manuscripts and printed books. But, there is no formal convention for usage of **𐤁**, and the letter is not used consistently across the sources. For example, not all words with /q/ are rendered using **𐤁** and in several records such words are written with **𐤀**. It is true that there are some sources in which **𐤁** is used for representing /k/, but these instances are idiosyncratic, and may be typographical errors. In L2/11-021 (p. 13), I showed the following entry from the Harvard catalogue (p. 283) to suggest that **𐤁** is a variant form of KA:








However, based on my analysis of this text, the gloss “**𐤁** = k” in the entry may be an error; it should refer to “qāf”. Throughout this text the **𐤁** is used for /q/, not /k/; except for one instance in which the letters are reversed for these values. The **𐤁** is used contrastively with **𐤀**, which is the regular form of /k/ in this text (see below for the example of *haqīqatī* from this very text). Moreover, in the numerous sources that I’ve consulted, when **𐤁** occurs in a document, it is always used concurrently with **𐤀**, which is in turn always the normative form of KA.

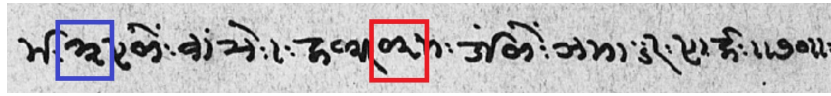
The usage of **𐤁** for /q/ is a significant innovation. It indicates a conscious effort by writers and printers of Khojki texts to contrast /q/ from /k/ not by using the typical combination of KA + NUKTA, but by using distinctive letters. If **𐤁** is palaeographically a stylistic variant of KA, Khoja scribes may have been inspired by the graphical similarity of the form to **𐤀** KA, such that they could assign to it a new value with phonetic similarity to U+n, in order to align similar letterforms with similar sound values. Given this similarity, it is quite possible that **𐤁** was intentionally created by scribes as a modification of **𐤀**. Its concurrent usage with **𐤀** established **𐤁** as a distinctive letter of the script with its own semantic identity.

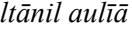
The **𐤁** is also notable as, to my knowledge, it is the only letter in any northern Indic script that is used distinctively for expressing /q/ independent of the letter for /k/. Apart from this letter, all other Arabic

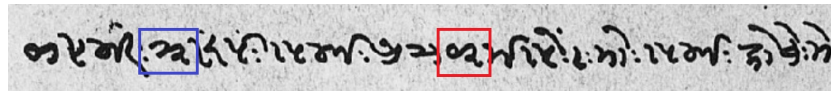
phonemes without Indic analogues, such as /ʕ/, /h/, /x/, /ɣ/, /z/, and /f/, are represented in Khojki using the sign NUKTA, eg.  'a,  ha,  ha,  ga,  za,  fa. Moreover, Khojki lacks a distinctive letter for the basic, high-frequency Arabic and Indic phoneme /ʃ/ = Brahmic SHA, which is graphically assimilated in the script with /s/ =  U+11229 KHOJKI LETTER SA, and represented using NUKTA, eg.  śa. Even SA + NUKTA does not uniquely express /ʃ/ as it is also used for the Arabic /θ/ and /sʃ/. Yet, Khojki users developed a distinctive method for writing /q/.


3 Examples of Usage

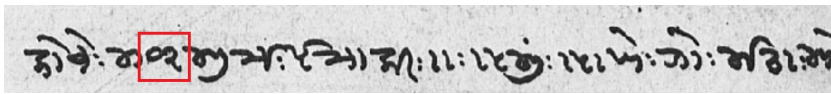
Regular usage of  (red) contrasted with  (blue) in manuscripts is shown below. For each excerpt, the name of the manuscript is given, as well as the Khojki word containing the letter *qa* and the corresponding Arabic term. Note: the visible, vowel-suppressing sign VIRAMA is conventionally indicated in these written records using  U+1122D KHOJKI VOWEL SIGN I or its variant form , depending upon the base letter, instead of the representative glyph  U+11235 KHOJKI SIGN VIRAMA.



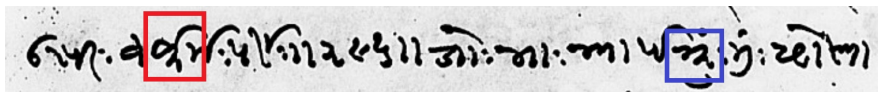
From *Kalāmi hazrat sultānil aulā* :  haqīqat = Arabic حقيقة haqīqat



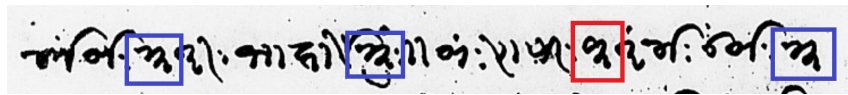
From *Kalāmi hazrat sultānil aulā* :  khalqat = Arabic خلقة khalqat




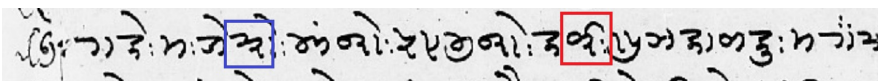
From *Kalāmi hazrat sultānil aulā* :  maqbul = Arabic مقبول maqbūl

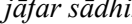


From *Miraj nama* :  vaqt = Arabic وقت waqt

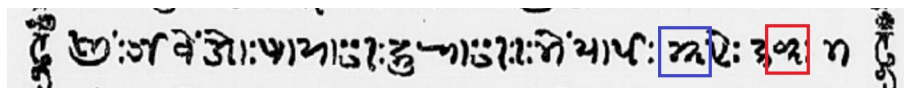


From *Mi'rāj nāma* :  qadam = Arabic قدم qadam

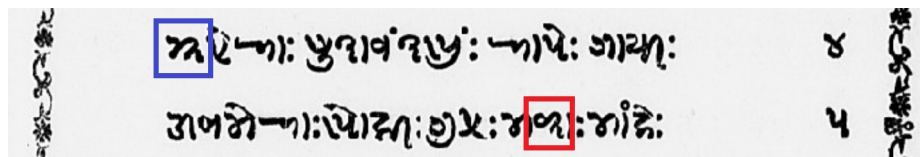


From *Rasālo hazrat emām jāfar sādhikjo* :  haq = Arabic حق haq

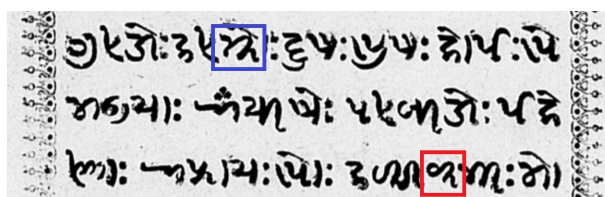
Regular usage of **𑂔** (red) contrasted with **𑂕** (blue) in printed documents is shown below. Here also, the name of the source is given, as well as the Khojki word containing the letter *qa* and the corresponding Arabic term.



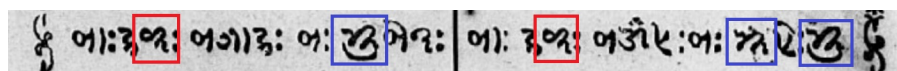
From *Elam sār saṅgrah rāg mālā* : 𑂕𑂔 *haq* = Arabic *حق haq*



From *Elam sār saṅgrah rāg mālā* : 𑂕𑂔𑂔[𑂔] *maqam* = Arabic *مقام maqam*

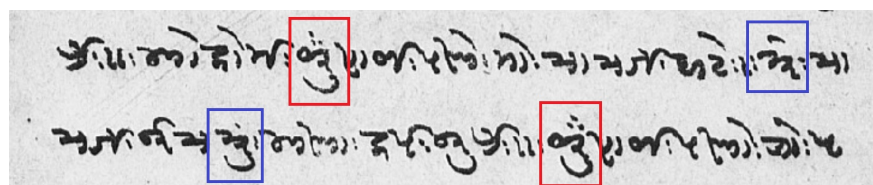


From *Ārtī dhūā vakhat trejī sāṅjījā ċogaḍīā 5* : 𑂕𑂔𑂔𑂔 *haqīqatī* = Arabic *حقيقتي haqīqatī*

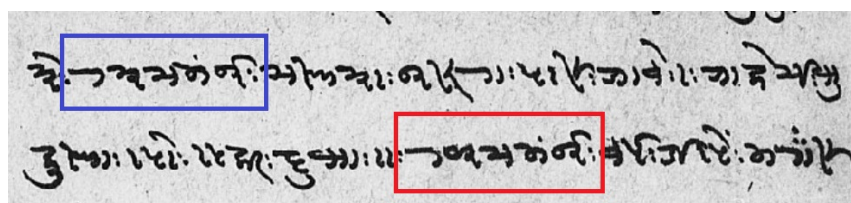


From *Phamdhāt zavāmardhī* : 𑂕𑂔 *haq* = Arabic *حق haq*

There are also irregular usages of **𑂔**. Shown below is contrastive usage of **𑂕** (blue) and **𑂔** (red), where **𑂔** written with NUKTA as **𑂔𑂔** in the word ‘Qurān’, the holy book of Islam. Such usage may be intended to emphasize the /q/ value of the letter, especially for words of significance to Islamic tradition.



The choice to use **𑂔** likely depends upon the scribe’s or printer’s awareness of the original spelling of Arabic words. The level of awareness can affect consistency usage of the letter, especially in handwritten documents. For example, a scribe may write a word with /q/ using **𑂕** *ka*, then when writing the word again, realize the /k/ is actually /q/. An example of such an orthographic ‘minimal pair’ occurs in the *Kalāmi hazrat sultānīl aulīā*, in the transliteration of the Persian term *عقل مند qī mand*, which is first written with /q/ simplified to /k/ as **𑂕𑂔𑂔𑂔**, then with preservation of /q/ in **𑂕𑂔𑂔𑂔**.



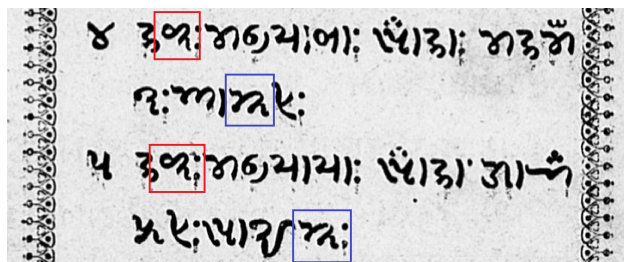
— ॐ नमो भगवते वासुदेवाय ॥ १ ॥ ॐ नमो भगवते वासुदेवाय ॥ १ ॥ ॐ नमो भगवते वासुदेवाय ॥ १ ॥

→ १४५६७८९१०१११२१३१४१५१६१७१८१९२०२१२२२३२४२५२६२७२८२९३०३१३२३३३४३५३६३७३८३९४०४१४२४३४४४५४६४७४८४९५०५१५२५३५४५५५६५७५८५९६०६१६२६३६४६५६६६७६८६९७०७१७२७३७४७५७६७७७८७९८०८१८२८३८४८५८६८७८८८९९०९१९२९३९४९५९६९७९८९९१००

જાયાજા: જૌમાળા: - જપ્પા: - (૧૦) જાળાળ:
 રૂઝાળો: ૫૦ ાળા
 જુમે: જાળાળાળ: (૨૩) જોજાળાળ: જાળાળાળ: ૧ મો:
 જાળાળાળ: પાજાળા:
 ————— ❁❁❁ —————
 જુમે: જાળાળાળ: જૌજાળાળ: જૌમાળાળ: **જાયાજા** જાળાળાળ

ॐ नमः शिवायः

There are cases where **𑂔** is used in high-frequency loanwords, such as **𑂔𑂔** *haq* = Arabic **حق** *ḥaq*, but not for other Arabic words containing /q/ that occur within the same line of a text, eg. **𑂔𑂔𑂔** *bakr* = Arabic **بقر** *baqr* and **𑂔𑂔𑂔𑂔** *sādika* = Arabic **صادق** *sādiq*, shown below. Here, *bākara* is part of a longer phrase *muhammad bākara* = Muḥammad al-Baqir and *sadika* is part of a longer phrase *jāaphara sādika* = Ja‘far al-Ṣādiq; these are the fifth and sixth *imām*-s of Shia Islam. The spelling of these names using **𑂔** instead of **𑂔** may be the conventional representation of the names of Shia *imām*-s in Khojki.



In all of the Khojki sources discussed above there is concurrent usage of **𑂔** and **𑂔**. Based upon the ‘regular’ usage of **𑂔** and **𑂔** for /q/ and /k/, it is evident that the writer or printer intended to use the two letters for different purposes. Even in ‘irregular’ cases where the sound values of the letters are switched, the letters are used contrastively. In printed texts, such discrepancies may be intentional and guided by the need to preserve the fidelity of the original manuscript upon which the text is based.

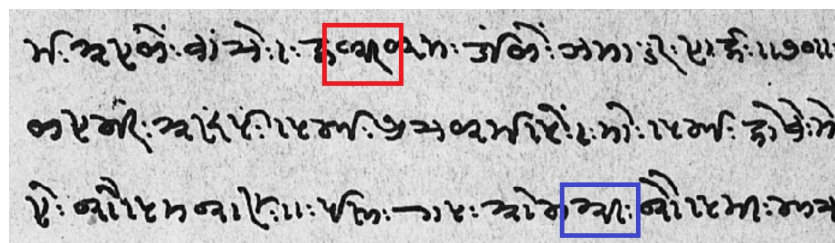
As Khojki is a liturgical script of Ismaili communities of South Asia, it is important that the Unicode repertoire for Khojki provide all characters required for the distinctive representation of records in digital plain text. To further enable such support, the **𑂔** KHOJKI LETTER QA should be encoded in the standard.

4 Collation

The **𑂔** KHOJKI LETTER QA should be sorted after **𑂔** U+11208 KHOJKI LETTER KA and before **𑂔** U+11209 KHOJKI LETTER KHA.

5 Glyph Interactions

The behavior of **𑂔** KHOJKI LETTER QA when combining with dependent vowel signs follows that for U+11208 KHOJKI LETTER KA. The excerpt below from From *Kalāmi hazrat sultānil aulīā* shows **𑂔** combined with **𑂔** U+1122D KHOJKI VOWEL SIGN II rendered as **𑂔𑂔** (red), as compared to the same vowel combination with **𑂔**, rendered as **𑂔𑂔** (blue):



6 Input Methods

On digital keyboards with touch-and-hold features, the 𑂔 KHOJKI LETTER QA should be added to the character palette for the 𑂔 U+11208 KHOJKI LETTER KA key. On other keyboards it may be placed on a SHIFT, CTRL, or ALT layer, mapped to the same key as U+11208 KHOJKI LETTER KA.

7 Character Data

Character Properties: UnicodeData.txt

1123F;KHOJKI LETTER QA;Lo;0;L;;;;;N;;;;;

Linebreaking Properties: LineBreak.txt

1123F;AL # Lo KHOJKI LETTER QA

Indic Syllabic Category: IndicSyllabicCategory.txt

1123F ; Consonant # Lo KHOJKI LETTER QA

8 References

Asani, Ali S. 1992. *The Harvard Collection of Ismaili Literature in Indic Languages: A Descriptive Catalog and Finding Aid*. Boston: G. K. Hall & Co.

Elam sār saṅgrah rāg mālā. Bombay: Lālībhaī Devrāj, Khojā Siṅdhā Čhāpākhānū, 1905. MS Indic 2534. Houghton Library, Harvard University, Cambridge, Mass.

Hazrat ‘Ali. Kalāmi:hazrat:sūltānil:auliā. Bombay: Alādhīn Gūlāmhūsen, 1878. MS Indic 2534. Houghton Library, Harvard University, Cambridge, Mass.

Ja‘far al-Šādiq. Rasālo:hazrat:emām:jāfar:sādhikjo. [s.l.: s.n., 1890]. MS Indic 2534. Houghton Library, Harvard University, Cambridge, Mass.

[Mi‘rāj nāma ?]. [s.l.; s.n.], 1849. MS Indic 2534. Houghton Library, Harvard University, Cambridge, Mass.

Pandey, Anshuman. 2011. “Final Proposal to Encode the Khojki Script in ISO/IEC 10646” (L2/11-021) <https://www.unicode.org/L2/L2011/11021-khojki.pdf>

Pīr Šadr ad-Dīn. Ārtī:dhūā:vakhat:tredjī:sāñjījā:ćogađā:5. Bombay: Khojā Lālībhaī Devrāj, Khojā Siṅdhā Čhāpākhānū, 1902. MS Indic 2534. Houghton Library, Harvard University, Cambridge, Mass.

Phaṇḍhīāt:zavāmardhī. Bombay: Lālībhaī Devrāj, Khojā Siṅdhā Čhāpākhānū, 1904. MS Indic 2534. Houghton Library, Harvard University, Cambridge, Mass.

**ISO/IEC JTC 1/SC 2/WG 2
PROPOSAL SUMMARY FORM TO ACCOMPANY SUBMISSIONS
FOR ADDITIONS TO THE REPERTOIRE OF ISO/IEC 10646¹**

Please fill all the sections A, B and C below.

Please read Principles and Procedures Document (P & P) from <http://std.dkuug.dk/JTC1/SC2/WG2/docs/principles.html> for guidelines and details before filling this form.

Please ensure you are using the latest Form from <http://std.dkuug.dk/JTC1/SC2/WG2/docs/summaryform.html>.

See also <http://std.dkuug.dk/JTC1/SC2/WG2/docs/roadmaps.html> for latest Roadmaps.

A. Administrative

1. Title:	Proposal to encode the Khojki Letter QA in Unicode		
2. Requester's name:	Anshuman Pandey <pandey@umich.edu>		
3. Requester type (Member body/Liaison/Individual contribution):	Expert contribution		
4. Submission date:	2021-05-21		
5. Requester's reference (if applicable):			
6. Choose one of the following:			
This is a complete proposal:			Yes
(or) More information will be provided later:			

B. Technical – General

1. Choose one of the following:			
a. This proposal is for a new script (set of characters):			
Proposed name of script:			
b. The proposal is for addition of character(s) to an existing block:			Yes
Name of the existing block:	Khojki		
2. Number of characters in proposal:			1
3. Proposed category (select one from below - see section 2.2 of P&P document):			
A-Contemporary <input checked="" type="checkbox"/>	B.1-Specialized (small collection)	B.2-Specialized (large collection)	
C-Major extinct <input type="checkbox"/>	D-Attested extinct	E-Minor extinct	
F-Archaic Hieroglyphic or Ideographic <input type="checkbox"/>	G-Obscure or questionable usage symbols		
4. Is a repertoire including character names provided?			Yes
a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document?			Yes
b. Are the character shapes attached in a legible form suitable for review?			Yes
5. Fonts related:			
a. Who will provide the appropriate computerized font to the Project Editor of 10646 for publishing the standard?	Anshuman Pandey		
b. Identify the party granting a license for use of the font by the editors (include address, e-mail, ftp-site, etc.):	Anshuman Pandey		
6. References:			
a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?			Yes
b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?	Yes		
7. Special encoding issues:			
Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)?			Yes

8. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at <http://www.unicode.org> for such information on other scripts. Also see Unicode Character Database (<http://www.unicode.org/reports/tr44/>) and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard.

¹ Form number: N4502-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11, 2005-01, 2005-09, 2005-10, 2007-03, 2008-05, 2009-11, 2011-03, 2012-01)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before?	No
If YES explain	
2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)?	No
If YES, with whom?	
If YES, available relevant documents:	
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?	Yes
Reference: See text of proposal	
4. The context of use for the proposed characters (type of use; common or rare)	Common
Reference: See text of proposal	
5. Are the proposed characters in current use by the user community?	Yes;
If YES, where? Reference: The Khoja Ismaili religious community and scholars	
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?	N/A
If YES, is a rationale provided?	
If YES, reference:	
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?	Yes
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?	No
If YES, is a rationale for its inclusion provided?	
If YES, reference:	
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?	No
If YES, is a rationale for its inclusion provided?	
If YES, reference:	
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character?	No
If YES, is a rationale for its inclusion provided?	
If YES, reference:	
11. Does the proposal include use of combining characters and/or use of composite sequences?	No
If YES, is a rationale for such use provided?	
If YES, reference:	
Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?	N/A
If YES, reference:	
12. Does the proposal contain characters with any special properties such as control function or similar semantics?	No
If YES, describe in detail (include attachment if necessary)	
13. Does the proposal contain any Ideographic compatibility characters?	No
If YES, are the equivalent corresponding unified ideographic characters identified?	
If YES, reference:	